Department of Economics

University of Oslo

Master's thesis

Modelling the effects of personality traits on ridership: The case of high speed rail in Norway

Author:
Bjørn Gjerde Johansen

Supervisor:
Prof. Erik Biørn

June 3, 2013

2013

Modeling the effects of personality traits on ridership:
The case of high speed rail in Norway

Bjørn Gjerde Johansen

http://www.duo.uio.no

Printed by Reprosentralen, Universitetet i Oslo

Acknowledgements

I am grateful to Erik Biørn for hours of discussion and supervision, including validation, suggestions for improvements and structural feedback; Institute of Transport Economics Norway and it's employees for access to the large and interesting dataset I am using, an office space with a coffee machine and motivational conversations throughout the semester; Stefan Flügel for providing me with the research topic, help with Biogeme and understanding the Biogeme output as well as giving me access to his unpublished articles and working papers about high speed rail in Norway; Farideh Ramjerdi for help understanding the program Biogeme and the necessary Python codes, help with writing the Python/Biogeme script for maximum likelihood estimation of the model parameters as well as discussions regarding the role of indicators for attitudinal, latent variables; and Eivind Hammersmark Olsen for proof-reading and structural feedback.

I would also like to thank James Odeck and the National Public Roads Administration for financial support.

Any errors or inaccuracies in this thesis are my own responsibility.

Executive Summary

A large-scale study of possibilities for and social benefits of high speed rail (HSR) in Norway has recently been conducted (Jernbaneverket, 2012). Following this, the subject of HSR has been frequently debated in Norwegian media. An important part of the cost-benefit analyses for HSR is the predicted ridership. Discrete choice modeling is the conventional method for estimating the mode choice probabilities used in these forecasts. Historically, the covariates taken into account in such models are attribute values for each modal choice as well as socio-economic attribute values for the travelers. However, even conditional on these variables there is often a high degree of individual, unobserved heterogeneity which contributes to low explanatory power. This is a potential problem, especially in the context of forecasting.

During the last decades, a lot of research has been done to better capture such individual heterogeneity. This thesis utilizes one of these methods described by Walker (2001) and Ben-Akiva et al. (2002) on the choice between air transport and HSR in Norway for business travelers. The method focuses on estimating the decision making process behind modal choice by including personality traits as latent variables in the utility functions.

These personality traits are mainly revealed through indicator variables in the form of questions regarding attitude and behaviors in everyday life. This can for instance be information regarding recycling behavior to reflect environmental consciousness, or information regarding safety behavior in traffic to reflect the preference for safety. The obvious advantage of such indicators is that information not inferable from market behavior can be included in the decision making process. If these latent variables are able to capture underlying personality traits, this may account for some of the unobserved heterogeneity and hence make forecasting more reliable.

In addition to reducing individual heterogeneity the model framework makes it possible to understand how different individual specific characteristics affect the personality traits. This allows for predicting different personality traits for different segments of individuals, and hence one should be able to predict the distribution of personality traits over the whole population. This is of particular interest in the context of forecasting.

My thesis consists of two parts. The first part is a complete analysis of the covariance structure of the indicators I have available. This consists mainly of exploratory and confirmatory factor analysis and results in suggestions for how personality traits best can be estimated based on these indicators. I provide suggestions for personality traits based on both of these methods and also establish the link between these personality traits and observable characteristics as income, gender and age.

The second part is integrated latent variable and choice models, where the personality traits “comfort” and “global environmental consciousness” are included as latent variables to explain the choice between air transport and HSR

in Norway. The market segment on which I focus is business travels on the links Oslo-Bergen and Oslo-Trondheim and the analysis is based on a stated preference study. I find that both these personality traits are significant. Moreover, they affect the choice probability for HSR positively and seem to do a better job in explaining mode choice than the available observable individual specific characteristics. I am cautious when drawing conclusions from the models since they are simple in terms of specification of utility functions. However, they shed light on aspects important for the utility of HSR that are easily forgotten in conventional analyses. This includes in particular the heterogeneity in how individuals' utilities are affected by changes in comfort, and the "purchase of moral satisfaction" by traveling more environmentally friendly.

Unfortunately, I am not able to show that individual heterogeneity is reduced in terms of increased explanatory power since I don't manage to provide a goodness of fit statistic for the estimated models. However, based on overall results and other similar case studies[1] I argue that the role of personality traits for the choice of HSR in Norway should be considered for further analysis; I have also outlined suggestions for how more sophisticated analyses can be conducted.

Finally, an important contribution of this thesis is that it summarizes the state of the art theories related to such analyses. It is to my knowledge no other sources in which theories regarding factor analyses, discrete choice models, latent variable models and a consistent framework in which latent variables enter the choice model are collected. In this manner my thesis provides added value for researches wanting to analyze choices in an attitudinal context since it describes the complete theoretical foundation of all the related processes.

A lesson learned worth to mention is that it is difficult to find observable variables that are good predictors of personality traits. Hence, a recommendation is that when designing a survey, care must be taken to figure out the relevant parts of the decision making process one wants to model as latent variables and also which observable attributes that may predict these latent variables.

[1] See for instance the three case studies described in Walker (2001), two case studies described in Ashok et al. (2002) as well as one case study in Johansson et al. (2006), one case study in Atasoy et al. (2010) and the case study related to latent variables in Morikawa (1989).

Contents

List of Tables

List of Figures

Abbreviations

CDF	Cumulative Distribution Function
CE1	Choice Experiment 1
CE2	Choice Experiment 2
CFA	Confirmatory Factor Analysis
CNL	Cross-Nested Logit
DC	Discrete Choice
D-M	Decision-Making
EFA	Exploratory factor analysis
FA	Factor Analysis
FIML	Full Information Maximum Likelihood
GEC	Global Environmental Consciousness
HSR	High Speed Rail
IIA	Independence of Irrelevant Alternatives
LEC	Local Environmental Consciousness
LISREL	Linear Structural Relationships system
LoS	Level of Service
MIMIC	Multiple Indicators, Multiple Causes
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
NL	Nested Logit
OLS	Ordinary Least Squares
PCA	Principal Components Analysis
PDF	Probability Density Function
RP	Revealed Preference
SEM	Structural Equation Model
SP	Stated Preference
TØI	Institute of Transport Economics
UIO	University of Oslo
VoT	Value of Time

1 Introduction

A large-scale study of possibilities for and social benefits of HSR in Norway has recently been conducted (Jernbaneverket, 2012), and following this the subject of HSR has been frequently debated in Norwegian media. The British consultant agency Atkins was hired to do the market analysis part of the study (Atkins, 2012a,b). For estimating the mode choice model, data from a stated preference study of binary choices between respondents' current mode of transport and HSR was used.

Another analysis is currently being conducted at Institute for Transport Economics Norway (TØI) based on a similar dataset (Flügel and Halse, 2012; Flügel et al., 2012). Here, some methodological weaknesses in the analysis by Atkins are pointed out. In particular, the contribution from TØI is a sophistication in terms of choice and specification of the discrete choice model. The analysis conducted by TØI is still ongoing, and in light of this which model that is most appropriate for estimating the demand for HSR in Norway still remains an open issue.

Discrete choice modeling is the conventional method for estimating mode choice probabilities used in forecasting models. Historically, the covariates taken into account in these models are attribute values for each modal choice as well as socio-economic attribute values for the travelers. This is also the case for both the Atkins and the TØI study. However, even conditional on these variables there is often a high degree of individual, unobserved heterogeneity which contributes to low explanatory power. This is a potential problem, especially in the context of forecasting. During the last decades, a lot of research has been done to better capture this individual heterogeneity. One method of doing this which also is the approach chosen for this thesis is by the use of unobservable latent variables based on indicator variables.

Typical indicator variables are questions of the form "*how important is it for you to ...*", or "*how often do you ...*", and respondents can for instance answer on a scale from one to five. The obvious advantage of these indicators is that information not inferable from market behavior can be included in the decision making process. This can for instance be information regarding recycling behavior to reflect environmental consciousness or information regarding safety behavior in traffic to reflect the preference for safety. If these latent variables are able to capture underlying personality traits, this may account for some of the unobserved heterogeneity and hence reduce uncertainty and make forecasting

more reliable.

This thesis will use the latter TØI dataset which contains such indicators. The focus will lie on including these indicator variables to predict latent variables that can be used directly in a choice model. These latent variables are assumed to reflect the personality traits "comfort", "flexibility", "reliability", "safety" and "global and local environmental consciousness".

In this thesis I will both describe a framework in which the effect of latent variables on the choice of HSR can be analyzed and also conduct an integrated latent variable and discrete choice analysis on the available dataset. This is not straight forward, and the two challenges that in my opinion are the gravest and which also are the main focus of the thesis will therefore hereby be described.

First, it is not always clear a priori which indicators that should be used to form which latent variables. Therefore, I utilize a number of different methods to better understand the covariance structure of the indicator variables and to see if it is reasonable to assume that the hypothesized latent constructs exist and whether their values are possible to predict based on the available data. This ranges from examining the correlation matrix to exploratory and confirmatory factor analysis. Exploratory factor analysis only uses the correlation structure of the indicator variables to create the proper latent factors. Confirmatory factor analysis is based on a priori assumptions regarding which latent variables that exist and the correlation structure between these latent variables and the indicators from the dataset. Ultimately, the results from this part of the thesis is used when formulating the latent variables for the choice model.

Second, including latent variables directly in the utility function will result in measurement errors in the choice model since the latent variables are observed with an error term. This leads to inconsistent estimators, and therefore a method for including these latent variables consistently must be used. The second part of the thesis is a description and an application of an integrated latent variable and choice model, developed and described by Walker (2001) and Ben-Akiva et al. (2002). This framework consists of (1) explicitly modeling the decision making process of the individual by the use of latent variables for different personality traits that are assumed to affect the preferences/utilities, and (2) including these variables in the choice model in a consistent and fully efficient way so that the whole model system can be estimated simultaneously by means of full information maximum likelihood. Figure 1.1 gives an overview of the assumed model structure[1]; squares represent observable variables and ellipses represent latent constructs. Arrows represent causal links. The boundaries of the latent variable model and the choice model are also indicated by brackets.

In addition to the application on the case study of hypothesized HSR in Norway, the two aforementioned main parts of this thesis contain a thorough description of the relevant theory through a summary of available state of the art literature.

[1]Figures representing the same relationships can be found in for instance Johansson et al. (2006); Walker (2001); Ben-Akiva et al. (2002); Atasoy et al. (2010).

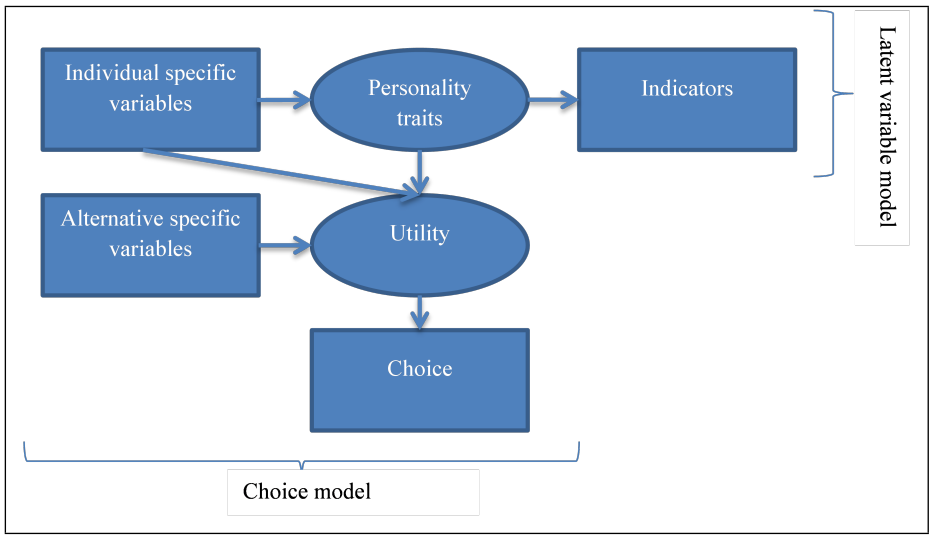

Figure 1.1: Integrated latent variable and choice model.

Even though the dataset used includes stated choices between both car, air, train or bus and HSR (see chapter 2 for information regarding the dataset), this thesis will only focus on the choice between HSR and air. This is done because of time constraints and hardware constraints[2]. For the same reasons, only the most simple form of the latent variable model is estimated. However, a range of possible extensions are suggested.

The thesis is organized as follows: Chapter 2 contains a description of the dataset. Chapter 3 synthesizes procedures regarding how to generate latent variables from indicators, where the main focus is the theory of factor analysis. It also includes applications of the methods resulting in factor analyses of the dataset. Finally, preliminary results for how the covariation of the indicator variables should be utilized to identify personality traits in a best possible way based on the factor analyses are included. Chapter 4 contains a theoretical framework for integrating latent variable models and discrete choice models. It also contains an application of this framework for the case of HSR versus air in Norway where models are estimated and discussed. Chapter 5 contains feasible and recommended extensions to the model estimated in chapter 4, hypothesized to give more realistic choice probabilities. Specifically the chapter contains extensions to the choice model and to the latent variable part of the model. Finally, chapter 6 contains concluding remarks.

It is important to emphasize that all theory described in this thesis are somebody else's work, and the relevant references are included throughout the

[2]For estimating the model simultaneously one has to (1) either optimize over an integral of a potentially high dimension which has to be solved numerically, or (2) simulate distributions from which a high number random draws are generated, which are both computationally demanding procedures.

document where it is appropriate. My own contribution only consists of summarizing these theories, as well as a simple application on the case of the demand for HSR in Norway.

Finally, this section will list the software I have used throughout the process. The thesis is written in LaTeX. For learning the Latex-language, the Latex Wikibook[3] has been of great help and is recommended to everyone wanting a quick tutorial or needing to check a particular command. All analyses in chapter 3 (EFA and CFA) are conducted in Stata 12 (StataCorp, 2011). Furthermore, Stata 12 is used for generating all tables of summary statistics and correlations that can be found throughout the thesis. For estimating the three models in chapter 4, the free and publicly available program Biogeme is used (Bierlaire, 2003). To estimate latent variable models in Biogeme, the new version that runs through Python must be used (Bierlaire and Fetiarison, 2009); this version allows for a more flexible specification of the likelihood function. To decide on the model specifications, preliminary analyses of the choice part and the latent variable part of the model were done separately in Stata. This was done for saving computation time.

[3]The LaTeXWikibook is constantly updated and can be accessed here: http://en.wikibooks.org/wiki/LaTeX.

2 Data

This thesis is based on a dataset from Institute of Transport Economics (TØI), who have conducted an independent survey to map the demand for high speed rail. This chapter will describe this dataset in detail. It is heavily based on a working paper from Institute of Transport Economics (Halse, 2012). If nothing else is stated, all information regarding the survey in general is collected from that paper. For more information regarding the survey, the reader is referred to appendix B as well as the aforementioned paper. Note that even though my analysis in chapter 4 is based on business travelers that chose between air and HSR, this chapter will describe the whole dataset.

2.1 Survey structure and choice experiments

This section describes the main features of the survey, which relates to the revealed preference (RP) and stated preference (SP) choices conducted by the respondents. They are described more in detail in appendix B. The survey consisted of two parts:

1. A RP survey where people (both business and leisure related) traveling by either car, plane, bus or train were stopped on the corridors Oslo-Trondheim and Oslo-Bergen and asked to fill in information regarding their trip. These corridors are the most relevant for a future high speed rail network within the boarders of Norway. The RP survey was a pen-and-pencil questionnaire, and originally a study of interregional travels in Norway (Denstadli and Gjerdåker, 2011). In total, about 8,500 respondents participated. Even though it is the dataset from the SP survey that is used in this thesis, the RP survey is relevant for two reasons: (1) the respondents from the SP survey are a subset of the RP survey respondents, and (2) the SP survey utilized characteristics from the "reference trip" from the RP survey as input.

2. A SP survey where respondents from the RP survey who had left their e-mail address (about 40 %) were contacted and asked to participate in a choice experiment. This survey was designed to reveal the ridership demand for high speed rail based on the RP survey data. Respondents were asked to state their preferred choice; either their reference trip or high

Table 2.1: Survey structure.

RP choices:							
Car		Rail		Bus		Plane	
SP choices:		SP choices:		SP choices:		SP choices:	
Car	HSR	Rail	HSR	Bus	HSR	Plane	HSR

speed rail at the same corridor. Hence, the main outputs from this survey are stated choices between car, plane, bus or rail and high speed rail. The variation in these choices arises from varying attribute values for the different alternatives. Participants made 14 SP choices each between HSR and the reference trip with varying attribute values for HSR; for choices 1–8 the attribute values for the reference trip were held constant, while for choices 9–14, the attribute values for the reference trip varied as well with a certain percentage below or above the reference value. In these last six choices, there were also a third alternative, *none*. The overall response rate was difficult to calculate since some e-mail addresses were corrected multiple times. However, it is assumed to be about 25 %. Considering this, and the fact that only 40 % of the RP study respondents left their e-mail addresses, there is clearly some selection bias present (Flügel, 2011).

This means that each respondent has made 15 choices; one RP choice between car, air, bus or train and 14 SP choices between the RP choice and HSR. In table 2.1 the structure of the data gathering (equivalent to the choice set of the respondents) is depicted. The six different attributes for both the reference alternative and HSR are (1) total cost, (2) in-vehicle time, (3) access time, (4) egress time, (5) frequency and (6) tunnel share (percentage of travel time in tunnel).

2.2 Behavioral and attitudinal indicators

In addition to the 14 different choice tasks, individuals responded to 23 questions regarding attitudes and personality traits. These questions will be the main focus of this thesis, and are displayed in table 2.2. The questions are based on a study from Sweden (Johansson et al., 2006). However, Johansson et al. applied the questions to short distance commuting trips, so some of the questions were irrelevant for high speed rail and were therefore changed. In this dataset there is also a separation between "local environmental consciousness" and "global environmental consciousness", since these factors are expected to affect the demand for high speed rail in opposite directions. Summary statistics for these indicators are displayed in table A.1 in appendix A. The correlation matrix for the indicator variables is displayed in table A.2. The same matrix is repeated in table 3.1, but here only correlations below -0.2 or above 0.2 are displayed so that it is easier to get an overview of the large correlations.

Table 2.2: Questions about attitudes and personality traits.

Question		Target dimensions
1	*How important is it for you* to be able to control the conditions around you (air condition, noise, music)?	Comfort
2	... to be able to rest on your trip?	
3	... to be able to work on your trip?	
4	... to avoid changing the mode of transport?	
5	... to know in advance how long the trip will take?	Reliability
6	... to have little or no variation in travel time?	
7	... to avoid congestion?	
8	... to have the opportunity to shop and make other errands?	Flexibility
9	... to be able to choose departure time yourself and be able to change it in short notice?	
10	... to have a car available at the destination?	
11	... to be able to choose travel route yourself and change it on the way?	
12	*How often do you* recycle batteries?	Local environmental consciousness
13s	... leave your garbage on the ground if there is no garbage can?	
14	... engage yourself to impede construction works and other activities that intervene nature?	
15	... visit unspoiled nature in order to experience it?	
16	... use a cycling helmet when you cycle?	Safety
17	... keep the speed limit when driving?	
18	... use the reflex when you walk in traffic in the dark?	
19s	... do things that are dangerous or illegal for fun?	
20s	... heat up your house so one does not have to use a sweater?	Global environmental consciousness
21	... turn off the lights before you leave the room?	
22	... bring shopping bags/used plastic bags when shopping?	
23	... do you eat dinner without meat?	

Note: Respondents answer with ordinal responses from 1 to 5 where 5 means *very important* or *always*. Questions 13, 19 and 20 are formulated with a negative meaning, and the scores are therefore "switched" ($1 = 5, 2 = 4, \dots$). To indicate this, they are marked with an *s*.

Questions 1–11 are *attitudinal* indicators, while questions 12–23 are *behavioural* indicators. Attitudinal indicators are of the form "how important is it for you…", while behavioral indicators are of the form "how often do you". Attitudinal indicators are meant to reflect attitudes that affect mode choice, while behavioral indicators are meant to represent behaviors that reflect attitudes that affect mode choice.

2.3 Descriptive statistics

Table 2.3 is a summary of the SP choices. In table 2.4, the variables from the dataset that are relevant for this thesis are described. In table 2.5, the variables for which it is appropriate have reported mean, standard deviation and number of observations.

Table 2.3: Summary of SP choices.

	RP choice				
SP choices	Car	Air	Train	Bus	Total
Reference mode	3,254	1,072	1,165	195	5,686
HSR	2,257	2,024	1,941	342	6,564
Neither	54	33	59	3	149
Total	5,565	3,129	3,165	540	12,399

From table 2.3 the distribution of SP choices conditional on the RP choice can be observed. It is apparent that car drivers have a relatively strong attitude towards the reference mode. This is not strange, considering that car is the mode of transport most unequal to HSR. This means that people with preferences against the attributes of HSR will be relatively more likely to choose car than other modes.

From table 2.5 it becomes apparent that in the choice experiments HSR was on average more expensive than the reference mode, and the in-vehicle time was on average shorter. Time to/from the station was on average longer and the average number of departures and the tunnel share were higher. The numbers are designed to be representative for an actual journey with the hypothesized mode HSR. One should note that the attributes *access/egress time* and *number of departures* are irrelevant for the mode car. 25 % of the trips are work trips, 36 % of the respondents are females and 13 % of the respondents brought a child on the reference trip. The average respondent is about 44 years old and has an income of about 450,000 NOK.

Table 2.4: List of relevant variables.

Choice variables	
valg	The SP choices between (1) reference mode, (2) HSR and (3) neither
trmiddel	The RP choice between (1) car, (2) air, (3) train and (4) bus
RP_SP_choice	The RP and SP choices between (1) car, (2) air, (3) train, (4) bus, (5) HSR and (6) neither
Alternative specific variables	
totkost_ref	Total cost (NOK) for the reference mode
totkost_hht	Total cost (NOK) for HSR
tidomb_ref	In-vehicle time (min) for the reference mode
tidomb_hht	In-vehicle time (min) for HSR
tidtil_ref	Access time (min) for the reference mode
tidtil_hht	Access time (min) for HSR
tidfra_ref	Egress time (min) for the reference mode
tidfra_hht	Egress time (min) for HSR
avg_ref	Number of departures per day for the reference mode
avg_hht	Number of departures per day for HSR
tunnel_ref	Share of the trip (%) inside a tunnel for the reference mode
tunnel_hht	Share of the trip (%) inside a tunnel for HSR
Individual specific variables	
age	Age of the respondent
income	Income of the respondent*
d_female	Dummy, = 1 if the respondent is a female
d_child	Dummy, = 1 if the respondent had a child below the age of 15 accompanying at the reference trip
d_worktrip	Dummy, = 1 if the reference trip was a work trip
Indicator variables 1–23	The behavioral and attitudinal indicator variables are displayed in figure 2.2

Note: All the variables are collected from the same source, namely the SP study which to some extent is based on the RP study.

* Respondents reported which income group they belonged to, and the average income of each group is then used for the income variable as an approximate value.

Table 2.5: Summary statistics of relevant variables.

Alternative specific variables			
Variable	Mean	S.D.	N
totkost_ref	698.605	497.82	12,399
totkost_hht	858.837	409.241	12,399
tidomb_ref	360.662	159.163	12,399
tidomb_hht	180.205	41.348	12,399
tidtil_ref	26.410	26.458	8,562
tidtil_hht	41.187	41.997	12,399
tidfra_ref	28.352	30.365	8,562
tidfra_hht	44.030	46.058	12,399
avg_ref	4.441	4.917	11,154
avg_hht	9.167	4.006	12,399
tunnel_ref	6.725	8.597	12,399
tunnel_hht	30.925	14.916	12,399
Individual specific variables			
Variable	Mean	S.D.	N
age	44.478	14.732	821
income	446,412.884	222,937.423	820
d_female	0.362	0.481	827
d_child	0.129	0.335	786
d_worktrip	0.258	0.438	827

3 Constructing latent variables from indicators

Before estimating a choice model with latent variables, one has to have a clear idea about how such latent variables should be formed. It is important to understand the theory behind how the questions[1] from table 2.2 will be transmitted into our choice model. Even though the questions are formulated based on hypothesized latent variables (the last column of table 2.2), it might be the case that the questions correlate in a different manner than first predicted. This chapter will present theories and methods for doing this, as well as ways of utilizing this theory to achieve preliminary results for the case of HSR in Norway.

Section 3.1 is meant to motivate the use of latent variables in the context of utility maximization and the use of indicator variables for estimating the latent variables. Section 3.2 contains the theory behind two different kinds of factor analysis (FA), exploratory and confirmatory. Factor analysis is a method for investigating whether a set of observable indicator variables are linearly related to another set of unobserved constructs with lower dimensionality, *factors*, and whether it is possible to generate factors that contain all the relevant variation in the observed variables[2].

Section 3.3 builds on the previous sections and investigates the indicators in this dataset by means of examining the correlation structure and conducting factor analyses. Finally, section 3.4 summarizes the results from section 3.3. This includes a preliminary conclusion for how the indicators should be used in a best possible way to construct the latent variables that are going to be used in the integrated choice and latent variable model in chapter 4.

[1] These questions will from now on be referred to as indicator variables.

[2] Another method for analyzing covariance structures by reducing the dimensionality of the data while at the same time preserve the maximal amount of variation is principal components analysis (PCA). This method is not appropriate for my dataset, since it does not take into account the latent structure of the hypothesized factors. To see why this is the case, the reader is referred to section C.2 in the theoretical annex for a brief description of PCA and section C.3 for a discussion on the pros and cons of PCA versus factor analysis. My motivations for including these sections are (1) that understanding PCA is crucial for understanding the "principal components approach" used for estimation in EFA, and (2) that PCA and factor analysis are often confused to be the same thing. I argue that it is important to be aware of both these methods, so that one is able to choose the appropriate one.

3.1 Motivation

This section is meant to motivate and discuss the use of indicators. Subsection 3.1.1 contains an overview of the decision making process and is meant to motivate the use of latent variables in decision processes. Subsection 3.1.2 contains a description and critical discussion of how these latent variables relate to indicator variables, and subsection 3.1.3 describes a potential problem when the indicators are meant to measure attitudes and not behaviors.

3.1.1 Decision making process

The sum of processes each individual goes through which lead from information to an actual choice is called the *decision making* (D-M) *process*, and only the outcome is observable to the researcher. The decision making process is therefore often referred to as a "black box". Ben-Akiva et al. (1999, p. 191) write

> "[. . .] the D-M process is defined as a sequence of mental operations used to transform the initial state of knowledge into a final *goal* state of knowledge."

Defining the processes in this black box explicitly should make the researcher better able to estimate realistic choice probabilities. The conventional utility maximization models incorporate the individual's preferences as latent variables, namely the *perceived utility* of each choice. The deterministic part of the utility function is then estimated based on available data. This is conventionally data that can be obtained from revealed or stated market behavior and observable socio-economic individual attributes.

However, the decision making process may well be too complex to be modeled through a direct link from observable attributes to utilities. A lot of research has been done recently in the cross section between econometrics and psychometrics to expand this black box to incorporate other latent variables. Examples of these are *attitudes*, *perceptions*, *motivation*, *memory* or *affect* (see for instance McFadden (1999); Ben-Akiva et al. (1999); Walker (2001) for more information regarding this). This thesis will focus on the inclusion of *attitudes* as latent variables in the same way as depicted in figure 1.1. The definition of attitudes as latent variables is adopted from McFadden (1999); Ben-Akiva et al. (2002). They write

> "*Attitudes* are defined as stable psychological tendencies to evaluate particular entities (outcomes or activities) with favor or disfavor".

Attitudes are further explained and the choice of attitudes at the most relevant latent variables is motivated by Ben-Akiva et al. (1999, p. 190). They write

> "Psychologists make a sharp distinction between attitudes and preferences. In this view, attitudes are multi-dimensional, with no requirement of consistency across attitudes. Preferences are viewed as constructed from more stable attitudes by a context-dependent

process that determines the prominence given to various attitudes and the trade-offs among them."

This indicates that including attitudinal variables in the choice process should increase both robustness and explanatory power. These attitudinal variables will be called *personality traits* in the remainder of this thesis, and will include for instance preferences toward comfort and environmental consciousness. Including these variables in the model system as in figure 1.1 should make a correctly specified model better able to capture individual heterogeneity in choice processes. To identify and estimate these personality traits, indicator variables obtained from questionnaires are used. These are described in the next section.

3.1.2 Relationship between indicators and personality traits

All indicator variables used in this thesis are displayed in table 2.2 in the chapter describing the dataset. This section is a motivation and critical discussion of whether such indicators can be used to capture personality traits. How indicators are used when personality traits are predicted is shown in figure 1.1, where the arrows indicate the direction of causality.

Intuitively, the use of indicator variables should extent our knowledge regarding individual behavior since it is a way to incorporate information that is not inferable from revealed or stated market behavior. Research has for instance indicated that a person with an *environmental* personality trait will perform more environmental behaviors than others (Ajzen and Fishbein, 1980, chapter 7), and therefore environmental behavioral indicators should be able to capture this personality trait.

However, including indicators will not necessarily improve a model. The first question one should ask in these kind of analyses is if the indicators really are able to capture the attitudes that affect modal choice. Lets continue the example of *environmental consciousness*, which also is a case that has received much attention in the literature. It may be the case that there is not complementarity between the indicators that are assumed to capture environmental behaviors and the choice of an environmental friendly mode. This is discussed by Johansson et al. (2006), and according to them it might happen for three main reasons:

- First, environmental friendly actions are more often performed when they are easy to perform. Actions that are perceived as costly or inconvenient cannot be expected to be performed even if the person displays an environmental friendly personality trait in other areas. The perceived costs of environmental behaviors may be highly heterogeneous. One could argue that all environmental indicators from this dataset (12–15 and 20–23) have a low cost relative to changing the mode of transport. Krantz Lindgren (2001) shows in interviews among individuals that drive regularly but still recognize the negative environmental effect of motorism that the perceived advantage of driving is relatively large compared to the perceived positive environmental effect of driving less.

- Second, environmental behaviors may be substitutes instead of complements. This may happen if users of a mode with relatively high environmental cost try to reduce their guilt in other areas of life. As an analogy Johansson et al. (2006) mention the term "risk compensation" from transport research; the overall perceived risk level is kept approximately constant, since drivers tend to increase speed when the road is perceived to be safe.

- Third, individuals may receive a "warm glow" from recycling (indicator 12), using less power for heating (indicator 20), turning off lights (indicator 21), bringing own shopping bags (indicator 22) and eating less meat (indicator 23). Warm glow is defined as the positive feeling of satisfaction one gets when doing something perceived to be good for the society. Kahneman and Knetsch (1992) call this "purchase of moral satisfaction". If choice of environmental friendly modes of transport do not give the same warm glow, it implies that transport and other environmental actions fulfill different needs, or are affected by different personality traits. See Andreoni (1989) for a formal analysis of this, in which he models giving charity and incorporates a warm glow effect.

This section has only contained examples of potential problems with indicators. However, it emphasizes that it is important to think thoroughly through whether the indicators actually capture the hypothesized personality trait or not.

3.1.3 Endogeneity of attitudinal indicators

As discussed in section 2.2 there are two kinds of indicators. "Attitudinal indicators" are responses to questions of the type "*how important is it for you to...*", while "behavioral indicators" are questions of the type "*how often do you...*". A potential problem with attitudinal indicators is that there might be a two-way causality between the level of the indicators and the individual's mode choice. This endogeneity problem is described by Morikawa (1989, p. 136):

> "[...] this hypothesis states that the respondent overstates the value of the psychometrics indicators of the chosen mode to justify his or her behavior, and, as a result, the perceptual indicators may contain information on the actual choice. This reversed relation of cause and effect is known as cognitive dissonance in psychology. Consequently, the latent variables which are linear combinations of the perceptual indicators have large explanatory power on the actual choice."

This potential problem may bias regressions, and according to Johansson et al. (2006) this is one reason to prefer behavioral indicators.

3.2 Theory

This section contains a summary of the theory needed to understand the methods which will be used to relate indicator variables to factors[3]. Factor analysis originates from psychometrics. The method dates back to the beginning of the 20th century and is generally ascribed to Charles Spearman. His earliest contribution relates results from a battery of psychological tests to a general "underlying, psychological" factor (Spearman, 1904) and he dedicated the rest of his professional life to develop and expand this method. Today however, factor analysis is found in most branches of statistical sciences.

There are two main types of factor analyses, *exploratory factor analysis* (EFA) where no assumptions are laid on the factorial structure of the data and factors are generated to best fit the observed variation, and *confirmatory factor analysis* (CFA) where the researcher has theoretical or empirical information a priori of the analysis, and this information is incorporated into the factor model by means of restrictions on model parameters. Subsection 3.2.1 contains the theory of exploratory factor analysis while subsection 3.2.2 contains the theory of confirmatory factor analysis. These sections are based on the book Rencher and Christensen (2012) unless stated otherwise.

3.2.1 Exploratory factor analysis

This section is a brief, formal description of EFA. It should be noted that one has to be familiar with eigenvalues and eigenvectors to completely understand the theory. See annex C.1 for a brief introduction to this subject. The framework used here is based on Rencher and Christensen (2012, chapter 13) unless stated otherwise, and all equations can be found more thoroughly described there.

Given p observable variables $y_1, y_2, \ldots, y_p$ (where individual specific subscripts are suppressed for simplicity) with mean values $\mu_1, \mu_2, \ldots, \mu_p$ and covariance matrix $\mathbf{\Sigma}$, we assume that the value of these variables are influenced by m unobservable, underlying *common factors*, $f_1, f_2, \ldots, f_m$ (where $m < p$) and an error term ϵ_i, in such a way that the underlying equation for the ith observable variable in the hypothesized model is

$$y_i - \mu_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \ldots + \lambda_{im} f_m + \epsilon_i \tag{3.1}$$

where λ_{ij} is the coefficient for how the ith variable relates to the jth factor[4]. These coefficients are called *factor loadings*. The system of p equations repre-

[3] "Factors" is the term used in factor analysis; these factors are represented by latent variables. I use the terms "latent variabls" and "factors" alternately, depending on whether I discuss factor analysis or latent variable models in general.

[4] In appedix C.2 I discuss principal components analysis (PCA). This is a method similar to EFA where the goal is to reduce a set of variables to a new set of variables with lower dimensionality while at the same time maintain the maximal amount of variation. The most obvious difference between EFA and PCA is that in PCA the unobserved "principal components" are modeled as constructs of the observed indicators, and not the other way around.

sented by equation 3.1 can also be written in matrix notation as

$$\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon} \tag{3.2}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$, $\mathbf{f} = (f_1, f_2, \ldots, f_m)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_p)'$ and $\boldsymbol{\Lambda}$ is a $(p \times m)$ matrix where the ijth element λ_{ij} is the coefficient for the jth factor from the ith equation.

It should be noted that it is possible to do EFA with both the correlation matrix and the covariance matrix as starting point[5]. The correlation matrix has correlations on the off-diagonal and units on the diagonal, while the covariance matrix has covariances on the off-diagonal and variances on the diagonal. If denoting correlation matrices by $\mathbf{R}$ and covariance matrices by $\boldsymbol{\Sigma}$, the relationship between these is that if $\mathbf{D} = \text{diag}(\boldsymbol{\Sigma})$, the diagonal matrix of variances, then $\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\boldsymbol{\Sigma}\mathbf{D}^{-\frac{1}{2}}$ so that the ijth correlation is the ijth covariance divided by the ith and the jth standard deviations. In the exposition below the method will be illustrated by use of the covariance matrix[6].

Assumptions

In EFA, some assumptions are imposed on the above variables. The standard assumptions imposed on f_j are zero expectation, $\text{E}(f_j) = 0, \forall j$, unit variance, $\text{var}(f_j) = 1, \forall j$ and zero covariance, $\text{cov}(f_j, f_k) = 0, j \neq k, \forall j, k$. In other words,

$$\text{E}(\mathbf{f}) = \mathbf{0} \tag{3.3}$$

$$\text{cov}(\mathbf{f}) = \mathbf{I}_m \tag{3.4}$$

where $\mathbf{I}_m$ denotes the $(m \times m)$ identity matrix. The assumptions for $\boldsymbol{\epsilon}$ are similar, but since ϵ_i is the residual part of y_i we have to allow for different variances. This gives the assumptions $\text{E}(\epsilon_i) = 0, \forall i$, $\text{var}(\epsilon_i) = \psi_i, \forall i$ and $\text{cov}(\epsilon_i, \epsilon_k) = 0, i \neq k, \forall i, k$. In addition, the regressors are assumed to be orthogonal to all of the error terms, $\text{cov}(\epsilon_i, f_j) = 0, \forall i, j$. These assumptions can be written as:

$$\text{E}(\boldsymbol{\epsilon}) = \mathbf{0} \tag{3.5}$$

$$\text{cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \text{diag}(\psi_1, \psi_2, \ldots, \psi_p) \tag{3.6}$$

$$\text{cov}(\mathbf{f}, \boldsymbol{\epsilon}) = \mathbf{0} \tag{3.7}$$

where "diag$(\cdot)$" denotes a matrix with the argument on the diagonal and zeros on the off-diagonals. Since all the factors have unit variance and are uncorrelated to each other and the error term, calculating the variance of y_i from equation 3.1 yields

$$\text{var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \ldots + \lambda_{im}^2 + \psi_i \tag{3.8}$$

In this expression $\lambda_{i1}^2 + \lambda_{i2}^2 + \ldots + \lambda_{im}^2 = h_i^2$ is called the common variance of variable i, or the *communality*, while ψ_i is called the specific variance of

[5] Correlations should be used if the variables are not measured in the same unit.

[6] It should be noted that the method I use in section 3.3.2 utilizes the correlation matrix by standardizing variables to unit variance. However, by using the conversion rule above, this method follows directly from the method described here.

variable i, or the *specificity*. The communality is the part of the variance of y_i explained by the factors, while the specificity is the unexplained part of the variance. Another property of the above model is that factor loadings represent covariances between factors and variables ($\text{cov}(y_i, f_j) = \lambda_{ij}, \forall i, j$, this follows from the previous assumptions), which can be written in matrix notation as:

$$\text{cov}(\mathbf{y}, \mathbf{f}) = \mathbf{\Lambda} \tag{3.9}$$

Estimation of factor loadings and specific variances

Using the EFA assumptions (equations 3.3–3.7) the covariance matrix of $\mathbf{y}$ can be written in terms of the factor loadings and the specific variance:

$$\begin{aligned}
\mathbf{\Sigma} &= \text{cov}(\mathbf{y}) \\
&= \text{cov}(\mathbf{\Lambda f} + \boldsymbol{\epsilon}) && \text{from relation 3.2} \\
&= \text{cov}(\mathbf{\Lambda f}) + \text{cov}(\boldsymbol{\epsilon}) && \text{by assumption 3.7} \\
&= \mathbf{\Lambda}\text{cov}(\mathbf{f})\mathbf{\Lambda}' + \mathbf{\Psi} && \text{by assumption 3.6} \\
&= \mathbf{\Lambda I \Lambda}' + \mathbf{\Psi} && \text{by assumption 3.4} \\
&= \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}
\end{aligned} \tag{3.10}$$

In the rest of this section I will describe the most common ways for estimating the factor loadings, starting with the most intuitive and ending with the most sophisticated[7].The standard way to estimate this expression is called the *principal component approach* (must not be confused with "principal components analysis", see appendix C.2), which will be explained below. One needs a random sample of n observations, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, to obtain the sample covariance matrix $\mathbf{S}$. Replacing the left hand side of equation 3.10 with $\mathbf{S}$ and the right hand side with the matrices' estimated counterparts, the new expression one seeks to estimate is $\mathbf{S} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$. The principal component approach focuses on $\hat{\mathbf{\Lambda}}$, and estimates $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$ first, independently of $\hat{\mathbf{\Psi}}$.

The first step is to eigen decompose $\mathbf{S}$ using normalized eigenvectors so that $\mathbf{S} = \mathbf{CDC}'$ where $\mathbf{D} = \text{diag}(\theta_1, \theta_2, \ldots, \theta_p)$ is the diagonal matrix of eigenvalues[8] and $\mathbf{C}$ is the orthogonal matrix of unit eigenvectors $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_p$ such that the ith column in $\mathbf{C}$, $\mathbf{c}_i = (c_{1i}, c_{2i}, \ldots, c_{pi})'$, is the eigenvector corresponding to the ith eigenvalue θ_i. This eigen decomposition can easily be derived from the theorem relating to equation C.3 in annex C.1 which describes the theory of eigenvectors and eigenvalues.

Since $\mathbf{D} = \mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}$ (this is always the case because $\mathbf{D}$ is always positive semidefinite) where $\mathbf{D}^{\frac{1}{2}} = \text{diag}(\sqrt{\theta_1}, \sqrt{\theta_2}, \ldots, \sqrt{\theta_p})$ it is possible to rewrite the empirical covariance matrix as $\mathbf{S} = \mathbf{CD}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{C}' = \mathbf{CD}^{\frac{1}{2}}(\mathbf{CD}^{\frac{1}{2}})'$. This form bears resemblance with the first term of the right hand side of equation 3.10; however, we need to reduce the matrix from $(p \times p)$ to $(p \times m)$. Therefore, two

[7] These are the three methods that were considered in section 3.3.2

[8] Following the notation of Rencher and Christensen (2012), I denote eigenvalues by θ instead of the standard notation λ to avoid confusion with factor loadings.

new matrices are defined: $\mathbf{D}_1 = \text{diag}(\theta_1, \theta_2, \ldots, \theta_m)$ and $\mathbf{C}_1 = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m)$ where the $p - m$ last (smallest) eigenvalues and the corresponding eigenvectors are removed. This is done since the eigenvalues represent how much variation that is contained in the variables, and by removing the rows and columns where the smallest variance of $\mathbf{S}$ is contained, one is able to reduce the dimensionality while minimizing the variation lost in the process[9]. Now, the estimators from equation 3.10 can be defined as

$$\begin{aligned}\hat{\mathbf{\Lambda}} &= \mathbf{C}_1\mathbf{D}_1^{\frac{1}{2}} = (\sqrt{\theta_1}\mathbf{c}_1, \sqrt{\theta_2}\mathbf{c}_2, \ldots, \sqrt{\theta_m}\mathbf{c}_m) \\ &= \begin{pmatrix} \sqrt{\theta_1}c_{11} & \sqrt{\theta_2}c_{12} & \ldots & \sqrt{\theta_m}c_{1m} \\ \sqrt{\theta_1}c_{21} & \sqrt{\theta_2}c_{22} & \ldots & \sqrt{\theta_m}c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\theta_1}c_{p1} & \sqrt{\theta_2}c_{p2} & \ldots & \sqrt{\theta_m}c_{pm} \end{pmatrix}\end{aligned} \tag{3.11}$$

and

$$\begin{aligned}\hat{\mathbf{\Psi}} &= \text{diag}(s_{11} - \sum_{j=1}^{m}\hat{\lambda}_{1j}^2, s_{22} - \sum_{j=1}^{m}\hat{\lambda}_{2j}^2, \ldots, s_{pp} - \sum_{j=1}^{m}\hat{\lambda}_{pj}^2) \\ &= \text{diag}(s_{11} - \hat{h}_1^2, s_{22} - \hat{h}_2^2, \ldots, s_{pp} - \hat{h}_p^2)\end{aligned} \tag{3.12}$$

where s_{ii} is the cell at the ith row and ith column of the covariance matrix. These are the principal component estimators of the factor loadings and the specific variances. $\hat{\mathbf{\Lambda}}$ is defined as to account for the variance resulting from the m first principal components, whereas $\hat{\mathbf{\Psi}}$ is defined as the residual variation in the diagonal terms. We notice therefore that the off-diagonal elements on the right hand side of $\mathbf{S} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$ are only approximately right, where the quality of the approximation depends on how much of the total variation that is contained in the m first eigenvectors (i.e. the relative size of the m first eigenvalues). The diagonal elements, however, are identical to the elements of the empirical covariance matrix because specific variances ψ_i are defined that way; they are added to account for the variation lost on the diagonal when removing the $p - m$ last eigenvectors.

Considering equation 3.11, it is worth noticing that the sum of squared cells of row i is equal to the ith communality, $\hat{h}_i^2$. This can easily be seen by the expression for communalities that is following equation 3.8 above. Furthermore, the sum of squared cells of column j is the jth eigenvalue of $\mathbf{S}$, $\sum_{i=1}^{p}(\sqrt{\theta_j}c_{ij})^2 = \theta_j\sum_{i=1}^{p}c_{ij}^2 = \theta_j$, because unit eigenvectors have a length of 1. The total sample variance is the sum of variances, in other words the trace of the covariance matrix, $s_{11} + s_{22} + \ldots + s_{pp} = \text{tr}(\mathbf{S})$. The part of the total variance that is due to the jth factor is therefore

$$\frac{\theta_j}{\text{tr}(\mathbf{S})} \tag{3.13}$$

[9]This is similar to the process of principal components analysis (PCA), which is described in section C.2 in the annex. See the discussion relating to equation C.5 in that section for more information on the relationship between eigenvalues and variation.

Another estimation method for factor loadings is the *principal factor method*. It uses an initial estimate $\hat{\boldsymbol{\Psi}}$ to obtain

$$\mathbf{S} - \hat{\boldsymbol{\Psi}} \approx \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' \tag{3.14}$$

so that the left hand side matrix is the covariance matrix, but with communalities instead of variances on the diagonal, and then estimates $\hat{\boldsymbol{\Lambda}}$ the same way as in equation 3.11. The relevance of both the principal factor method and the principal component method lies in estimating $\hat{\boldsymbol{\Lambda}}$. The advantage of the principal factor method, however, is that the specificities are taken into account when the factor loadings are estimated. It is conventional to use variance scaled by the squared multiple correlation between y_i and the other $p-1$ variables as an initial estimate of the ith communality, $\hat{h}_i^2 = s_{ii}R_i^2$.

Another way of estimating $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ is by maximum likelihood (ML). This is perhaps the most obvious way of estimation because of ML's intuitive appeal; however, it requires a strict assumption about multivariate normality. If we assume that the observations $y_1, y_2, \ldots, y_n$ constitute a random sample from $\mathcal{N}_p(\mu, \boldsymbol{\Sigma})$, then it can be shown (Rencher and Christensen, 2012, chapter 13) that the following

$$\mathbf{S}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\Lambda}}(\mathbf{I} + \hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\boldsymbol{\Lambda}) \tag{3.15}$$

$$\hat{\boldsymbol{\Psi}} = \operatorname{diag}(\mathbf{S} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}') \tag{3.16}$$

$$\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} \qquad \text{is diagonal} \tag{3.17}$$

has to be satisfied for the estimates $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$. Solving the equations will therefore yield the maximum likelihood estimates.

Rotation

An important property of the factor loadings is that they are not unique. In fact, $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}$, where $\mathbf{T}$ is any orthogonal matrix, will reproduce the same covariance matrix as $\boldsymbol{\Lambda}$ does. See Rencher and Christensen (2012, chapter 13, p. 441–442) for a formal proof of this. Such transformations are called *orthogonal rotations* since multiplication with orthogonal matrices is the same as rotating the axes (angles, distances and communalities remain unchanged). This is similar to PCA (see section C.2 in the appendix), which can be viewed as a rotation around the multidimensional mean. Unlike PCA however, the observed indicators in EFA are not affected by the rotation. This is because the loadings are applied to the factors, which are only underlying constructs of the observed variables.

Such rotations are convenient when interpreting the factor loadings; if the axes are rotated in such a way that points lie close to an axis (or more axes), the observations load highly on the factor(s) corresponding to that axis (those axes). By examining which variables each factor is affecting and in which direction, the factors can be interpreted.

The most popular method for orthogonal rotations is called the *varimax* technique. This technique finds an orthogonal matrix $\mathbf{T}$ that maximizes the variance of the squared loadings in each column of $\hat{\mathbf{\Lambda}}^*$. The maximum variance is obtained when some loadings are as large (in absolute numbers) as possible while other loadings are as close to zero as possible. This makes the factors easy to interpret because each factor will influence some indicators greatly, while other indicators will not be influenced at all. The opposite of this is if all loadings in a column are nearly equal; then the variance would approach zero and indicators would be influenced equally much by all factors.

There is also something called an *oblique rotation*, which is a non-orthogonal transformation. The misleading term oblique *rotation* is well established in the literature; however, an oblique *transformation* would be a more accurate description since non-orthogonal transformations do not preserve distances and hence they are more than rotations of the axes. Oblique rotations alter distances, angles and communalities and lead to new factors that are correlated. The advantage is that since axes are not restricted to be perpendicular, oblique axes are often able to pass closer to the observations; however, more care has to be taken when it comes to interpretation of the rotated/transformed factors.

Estimation of factor scores

Estimating the factor scores $\hat{\mathbf{f}}_i = (\hat{f}_{i1}, \hat{f}_{i2}, \ldots, \hat{f}_{im})'$ for each of the n observations is not necessary if one only wants an overview of the covariance structure of the data; however, if one plans to use the factors for further analysis these have to be estimated as well. The most usual way of estimating the factor scores is by regression. Since the mean of each factor is assumed to be zero, the jth of the m factors is modeled as (suppressing the individual specific subscript)

$$f_j = \beta_{j1}(y_1 - \bar{y}_1) + \beta_{j2}(y_2 - \bar{y}_2) + \ldots + \beta_{jp}(y_p - \bar{y}_p) + \xi_j \tag{3.18}$$

so that the system of m equations for individual i becomes

$$\mathbf{f}_i = \mathbf{B}_1'(\mathbf{y}_i - \bar{\mathbf{y}}) + \boldsymbol{\xi}_i \quad i = 1, 2, \ldots, n \tag{3.19}$$

where the vector $\mathbf{f}_i$ is $(m \times 1)$, the vector $(\mathbf{y}_i - \bar{\mathbf{y}})$ is $(p \times 1)$, the vector $\boldsymbol{\xi}_i$ is $(m \times 1)$ and $\mathbf{B}_1'$ is a $(m \times p)$ matrix of coefficients with no intercept. Using the transposed form which is $\mathbf{f}_i' = (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{B}_1 + \boldsymbol{\xi}_i'$, these n equation systems can be combined to one model:

$$\begin{aligned}\mathbf{F} = \begin{pmatrix}\mathbf{f}_1' \\ \mathbf{f}_2' \\ \vdots \\ \mathbf{f}_n'\end{pmatrix} &= \begin{pmatrix}(\mathbf{y}_1 - \bar{\mathbf{y}})' \\ (\mathbf{y}_2 - \bar{\mathbf{y}})' \\ \vdots \\ (\mathbf{y}_n - \bar{\mathbf{y}})'\end{pmatrix}\mathbf{B}_1 + \begin{pmatrix}\boldsymbol{\xi}_1' \\ \boldsymbol{\xi}_2' \\ \vdots \\ \boldsymbol{\xi}_n'\end{pmatrix} \\ &= \mathbf{Y}_c\mathbf{B}_1 + \mathbf{\Xi}\end{aligned} \tag{3.20}$$

where $\mathbf{Y}_c$ denotes centered (mean-reduced) variables and $\mathbf{\Xi}$ is the error matrix. The conventional estimator for this type of coefficient matrix is $\hat{\mathbf{B}}_1 =$

$(\mathbf{Y}_c'\mathbf{Y}_c)^{-1}\mathbf{Y}_c'\mathbf{F}$. This is not a feasible estimator in factor analysis since $\mathbf{F}$ is unobservable. However, by multiplying and dividing the expression by $(n-1)$ we obtain (Rencher and Christensen, 2012, p. 362):

$$\hat{\mathbf{B}}_1 = \left(\frac{\mathbf{Y}_c'\mathbf{Y}_c}{n-1}\right)^{-1} \frac{\mathbf{Y}_c'\mathbf{F}}{n-1} = \mathbf{S}_{yy}^{-1}\mathbf{S}_{yf} \tag{3.21}$$

$\mathbf{S}_{yy}^{-1}$ denotes the inverse of the covariance matrix for which the simplified notation $\mathbf{S}$ is previously used, and looking at equation 3.9, we see that $\hat{\mathbf{\Lambda}}$ is an estimate of the covariances between $\mathbf{y}$ and $\mathbf{f}$, $\mathbf{S}_{yf}$. Hence, the first term of the right hand side of the last equality sign is observed while the last term has an estimated counterpart, and $\hat{\mathbf{B}}_1 = \mathbf{S}^{-1}\hat{\mathbf{\Lambda}}$ is therefore a feasible estimator of $\mathbf{B}_1$. Inserting this in equation 3.20, the predicted values for all factors for all observation pairs in matrix notation are

$$\hat{\mathbf{F}} = \mathbf{Y}_c\mathbf{S}^{-1}\hat{\mathbf{\Lambda}} \tag{3.22}$$

3.2.2 Confirmatory factor analysis

This exposition is based on Rencher and Christensen (2012, chapter 14). Confirmatory factor analysis (CFA) is different from EFA in that it incorporates constraints into the model; EFA is applied when the factorial structure is unknown, while CFA is used in situations when one has some (theoretical or empirical) information beforehand regarding the structure of the model parameters. CFA is often defined within the subject of structural equation models (SEM). The factor analysis model is very similar to that described in section 3.2.1. Given p observable variables $y_1, y_2, \ldots, y_p$ (where individual specific subscripts are suppressed for simplicity) with mean values $\mu_1, \mu_2, \ldots, \mu_p$ and covariance matrix $\mathbf{\Sigma}$, we assume that the value of these variables are influenced by m unobservable, underlying factors, $\eta_1, \eta_2, \ldots, \eta_m$[10] (where $m < p$) and an error term ϵ_i, in such a way that the underlying equation for the ith observable variable is

$$y_i = \mu_i + \lambda_{i1}\eta_1 + \lambda_{i2}\eta_2 + \ldots + \lambda_{im}\eta_m + \epsilon_i \tag{3.23}$$

where λ_{ij} represent the structural coefficient for indicator i and factor j, which is equivalent to equation 3.1. In matrix notation, the expression becomes

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{3.24}$$

which is equivalent to equation 3.2. The error vector is defined similar as for EFA; however, the factors are usually allowed to be correlated with each other. In other words, both vectors $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_p)'$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)'$ have mean $\mathbf{0}$ and covariances

$$\text{cov}(\boldsymbol{\epsilon}) = \mathbf{\Psi} = \text{diag}(\psi_{11}, \psi_{22}, \ldots, \psi_{pp}) \tag{3.25}$$

[10] To separate EFA and CFA, factors from EFA are referred to as f_i while factors from CFA are referred to as η_i.

and

$$\text{cov}(\boldsymbol{\eta}) = \boldsymbol{\Phi} = \begin{pmatrix} \phi_{11} & \phi_{12} & \dots & \phi_{1m} \\ \phi_{21} & \phi_{22} & \dots & \phi_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{m1} & \phi_{m2} & \dots & \phi_{mm} \end{pmatrix} \tag{3.26}$$

Equation 3.25 implies that there can be no correlation between the indicators, except through the assumed factors. As for EFA, this model implies a specific correlation structure so that the covariance matrix can be defined as

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \tag{3.27}$$

when the model holds, where $\boldsymbol{\theta} = (\boldsymbol{\lambda}', \boldsymbol{\phi}', \boldsymbol{\psi}')'$ is a vector of the model parameters. For a model that is unconstrained, $\boldsymbol{\lambda}$ contains pm factor loadings, $\boldsymbol{\phi}$ contains $m(m+1)/2$ factor variances and covariances and $\boldsymbol{\psi}$ contains p error variances. The order condition is satisfied if the number of variances and covariances from the covariance matrix is at least as large as the number of specified parameters to be estimated in the model. The covariance matrix has $p(p+1)/2$ unique cells, and therefore the order condition is

$$\frac{p(p+1)}{2} \geq pm + \frac{m(m+1)}{2} + p \tag{3.28}$$

where the right hand side of the equation is the dimension of the parameter vector $\boldsymbol{\theta}$. A popular way of constraining the model to ensure that the order condition is satisfied is by the following setup:

$$\begin{pmatrix} \mathbf{y}_0 \\ m \times 1 \\ \mathbf{y}_1 \\ (p-m) \times 1 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ m \times 1 \\ \boldsymbol{\mu}_1 \\ (p-m) \times 1 \end{pmatrix} + \begin{pmatrix} \mathbf{I} \\ m \times m \\ \boldsymbol{\Lambda}_1 \\ (p-m) \times m \end{pmatrix} \boldsymbol{\eta} + \begin{pmatrix} \boldsymbol{\epsilon}_0 \\ m \times 1 \\ \boldsymbol{\epsilon}_1 \\ (p-m) \times 1 \end{pmatrix} \tag{3.29}$$

Here, the m factors are set equal to one indicator each, so that the equations for m of the indicators contained in $\mathbf{y}$ are constrained to $y_i = f_i + \epsilon_i, \quad i \in [1, m]$. This model is however only an example that satisfies the order restriction; other restrictions on the parameters should be imposed if they are consistent with theory and a priori expectations about how factors relate to the indicators.

Estimation of the parameters is based on minimizing the difference between $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\mathbf{S}$, where $\boldsymbol{\theta}$ can take values within the parameter space $\boldsymbol{\Theta}$, which is defined by the researcher based on parameter restrictions consistent with theory and a priori expectations. Based on joint normality assumptions, maximum likelihood estimators can be derived. Furthermore, it has been shown that these estimators are consistent for non-normal samples as well, as long as the number of observations is sufficiently large (Amemiya and Anderson, 1990; Anderson and Amemiya, 1988). It can be shown (Rencher and Christensen, 2012, chapter 14, p. 487) that the likelihood for $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ given $\mathbf{S}_n$, where $\mathbf{S}_n$ is the covariance matrix given that $\mathbf{y}$ is multivariate normal distributed, $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$L[\boldsymbol{\Sigma}(\boldsymbol{\theta}); \mathbf{S}_n] = c \times |\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{\frac{-n}{2}} \exp\Big(-\frac{1}{2}\text{tr}\big\{ n\mathbf{S}_n[\boldsymbol{\Sigma}(\boldsymbol{\theta})]^{-1} \big\} \Big) \tag{3.30}$$

where the maximum likelihood estimator for $\boldsymbol{\theta}$, denoted $\hat{\boldsymbol{\theta}}_{ML}$, will ensure that $\hat{\boldsymbol{\lambda}}_{ML}$ converge to $\boldsymbol{\lambda}$ when $n \to \infty$, given the defined parameter space $\boldsymbol{\Theta}$.

3.3 Application

This section builds on the previous theory, and includes various analyses of the indicator variables in order to expand our knowledge on how they should be used to form latent variables. Subsection 3.3.1 concentrates on the correlation matrix, in subsection 3.3.2 exploratory factor analysis is conducted and in subsection 3.3.3 confirmatory factor analysis is conducted.

3.3.1 Examining the correlation matrix

This subsection will focus on the correlation matrix in table 3.1. Here, values below 0.2 in absolute value are removed to get a better overview of the large correlations. The same table with all correlations included can be found in the appendix for completeness's sake (table A.2). Looking at table 3.1 in relation to the target dimensions from table 2.2, we see that most of the indicators within each target dimension have a positive, significant correlation. The highest correlations within each category can be found for the target dimensions reliability and flexibility. Furthermore, we see that (1) reliability indicators are highly correlated with comfort indicators, and (2) local environmental consciousness indicators are correlated with global environmental consciousness indicators.

Regarding (1), it could be because the dimensions attract particular groups of people that are more homogeneous; families with small children will for instance often require both reliability and comfort, and be less interested in flexibility since more planning often is required in advance when they are traveling with their children.

Regarding (2), even though the correlations between global and local environmental consciousness indicators are high, it is not obvious that they can be merged. This is because they are hypothesized to affect the demand for high speed rail in opposite directions. On the other hand, even though HSR affects the local environment to the worse, people may consider that as a sunk cost. This is equivalent to saying that if HSR exists, peoples' preference for the local environment will not affect their use of it because it is the construction and the ridership possibility for others (through e.g. housing expansion) and not the direct use of HSR for the respondent that influences the local environment.

Table 3.1 reveals that all indicator pairs with a correlation lower than -0.2 contain indicator 10, and these question pairs will therefore be examined. Question 10 belongs to the target dimension flexibility and reads *how important is it for you to have a car available at the destination?* This is obviously a question relating to car drivers in particular. The negative correlations result from questions 2, 3 and 5, of which the two first belong to the target dimension comfort, and the latter to reliability.

Table 3.1: Correlation matrix of behavioral and attitudinal indicators, small values are not displayed.

$N = 502$	Indicators:																		
Indicators:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1.000																		
2		1.000																	
3		0.490	1.000																
4	0.264			1.000															
5	0.204	0.493	0.379	0.322	1.000														
6		0.297	0.307	0.263	0.471	1.000													
7	0.217	0.282	0.239	0.355	0.452	0.364	1.000												
8								1.000											
9	0.211			0.273		0.234		0.218	1.000										
10		-0.256	-0.210		-0.234			0.270	0.252	1.000									
11	0.233							0.326	0.506	0.391	1.000								
12												1.000							
13												0.205	1.000						
14														1.000					
15												0.235		0.302	1.000				
16												0.210				1.000			
17																	1.000		
18												0.268				0.286	0.221	1.000	
19																	0.303	0.256	1.000
20																			
21																			
22												0.317			0.254		0.210		
23												0.225							

	Indicators:			
Indicators:	20	21	22	23
20	1.000			
21		1.000		
22	0.207	0.208	1.000	
23				1.000

Note: The 23 indicators of which the correlations are displayed in this table are the same as the 23 questions regarding behaviors and attitudes. It is assumed that indicators 1–4 relate to *comfort*, indicators 5–7 relate to *reliability*, indicators 8–11 relate to *flexibility*, indicators 12–15 relate to *local environmental consciousness*, indicators 16–19 relate to *safety* while indicators 20–23 relate to *global environmental consciousness*. Horizontal and vertical lines are added to group these indicators together. The questions that are formulated with a negative meaning are reversed, so that the correlations are meaningful. Values below 0.2 in absolute value are not displayed.

Table 3.2: Indicator variables 2, 3 and 5 in relation to driving a car.

Indicator variables:	2		3		5	
	Mean:	N:	Mean:	N:	Mean:	N:
Do not own car:	4.46	147	3.71	147	4.48	147
	(0.71)		(1.07)		(0.62)	
Own car:	3.82	680	3.05	680	4.12	680
	(1.03)		(1.26)		(0.91)	
Used car at	3.48	371	2.69	371	3.92	371
reference trip:	(1.06)		(1.19)		(0.99)	

Note: Participants of the survey have answered from 1 to 5, which denotes not important to very important respectively. The mean of this is only meaningful if one assumes a linear relationship between the five alternatives (so that the answer 4 means twice as important as the answer 2).

Examining table 2.2, we see that questions 2 and 3 are concerned with being able to rest or work on a trip, which is of course impossible when one drives. Question 5 reads *how important is it for you to know in advance how long the trip will take?* This relates to all modes, but one should expect different perceptions of the question for, say, car drivers and bus users; even with the same value of time (VoT) and the same risk perception they may respond differently to the question since car drivers perceive the concept of delays as relatively short (only in terms of traffic variation) while bus users perceive the concept of delays as relatively long (traffic variation plus variation in departure time). Therefore, it is logical that the extent to which one reports to care about the variation of the trip is negatively correlated with the dummy for being a car driver.

This is also indicated in table 3.2, where the mean scores of the indicators 2, 3 and 5 are calculated for the sub-populations that (1) do not own a car, (2) have reported to own a car and (3) own a car and chose car as mode on the reference trip. As expected, the mean scores are decreasing for all three indicators[11].

These phenomena are examples of reversed causality; it is not only the factors that affect choice of mode, but choice of mode also affects the factors, through indicator scores. This potential problem was also mentioned in section 3.1.3.

3.3.2 Exploratory factor analysis

In this section the EFA performed on the data is described. Flügel (2011) also performs an EFA on the same dataset. My further analysis will be based on the joint findings from this EFA and the EFA described by Flügel. I use the computer program Stata 12 (StataCorp, 2011) for all operations.

[11] The mean scores are not cardinally meaningful before one assumes a linear relationship between the five alternatives (so that the answer 4 means twice as important as the answer 2). However, they are still informative as ordinal values.

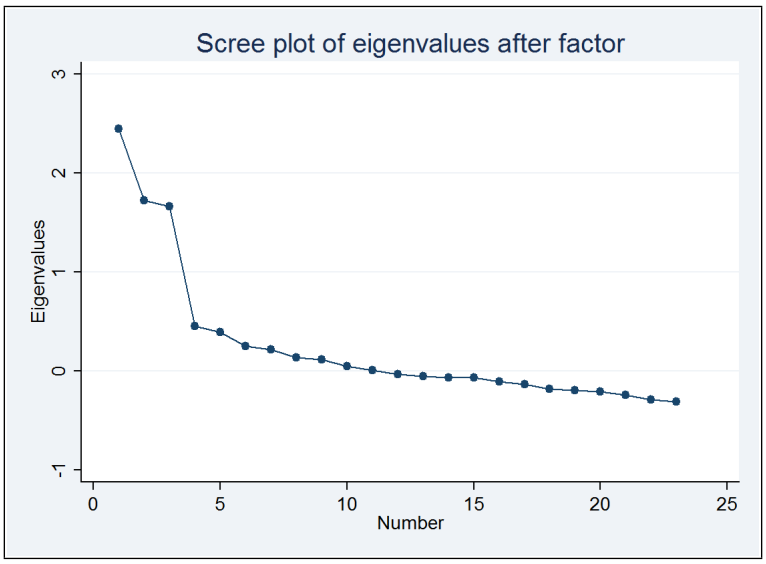

Figure 3.1: Scree plot after exploratory factor analysis, displaying all 23 eigen values.

The fact that the indicators are far from normally distributed makes maximum likelihood estimation of factor loadings and specificities inappropriate. Since the specificities are relatively high, a method that takes these into account is preferable. The principal component method does not do this, and is therefore not optimal. The principal factor method does not require joint normality and takes specificities into account, and is therefore the estimation method chosen (a description of the principal factor method is found in relation to equation 3.13).

In factor analysis, the number of factors to retain is an important consideration. Unlike PCA, adding or removing a factor will alter the loadings for all other factors as well. The scree test is a test for choosing the number of factors to retain. After an unrestricted EFA, the eigenvalues for all the p factors should be plotted. This is done in figure 3.1 for all 23 eigenvalues. The scree test states that if the graph drops sharply, followed by a straight line with much smaller slope, choose m equal to the number of eigenvalues before the straight line begins (Rencher and Christensen, 2012, chapter 13). According to figure 3.1, the number of factors to be retained is three.

Another criterion for how to choose the number of factors is, "choose m equal to the number of eigenvalues greater than the average eigenvalue" (Rencher and Christensen, 2012, chapter 13, p. 453)[12]. This criterion also suggests that three factors should be kept, and this further increases our confidence in the choice

[12]The average eigenvalue (the average variance) is of course 1 if one uses the correlation matrix for the eigen decomposition, because before any factors are discarded the factors contain the exact same variation as the original indicators.

of retaining three factors.

According to the retainment criteria in Flügel (2011, p. 44), seven factors should be kept[13]. When different criteria lead to different number of factors, the validity of the results are questionable. This is particularly the case with EFA, since it is exposed to a great deal of subjectivity. However, another important criterion for choosing the number of factors is interpretability. Flügel showed that a high number of factors makes each individual factor difficult to interpret. The next sections show that with three factors instead of seven, the interpretability is greatly improved.

Restricting the number of factors to three, re-estimating the factor loadings and rotating them with the orthogonal varimax rotation gives the factor loadings and specificities shown in table 3.3. In this table values below 0.3 in absolute value are shown as blanks to lay emphasis on the largest correlations. For the sake of completeness, the same table where all loadings are displayed is included in the appendix as table A.3

The estimated specificities $(\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_{23})$ are shown in the last column. The communalities are found by subtracting the specificity from one: $\hat{h}_i^2 = 1 - \hat{\psi}_i, \quad i \in [1, 23]$. The loadings associated with each of the factors (for the moment only named as factor 1, 2 and 3) are displayed in the middle columns. These columns constitute the (23×3) matrix $\hat{\mathbf{\Lambda}}$.

One sees immediately that the attitudinal indicators are explained by factors 1 and 3 while the behavioral indicators are explained by factor 2. Factor 1 is similar to the first factor in Flügel (2011) and may be interpreted as "convenience". The loadings are high for all indicators associated with comfort and reliability. The only indicator with a factor loading below 0.3 (the factor loading is 0.29) is indicator 1, *how important is it for you to be able to control the conditions around you?*[14] Factor 2 may be interpreted as "carefulness" or "political correctness" and it measures the degree of which one takes the consequences of one's actions into account. All indicators relating to local and global environmental consciousness and safety load highest on this factor. Factor 3 may be interpreted as "flexibility"; all indicators with the target dimension flexibility load highest on this factor, as well as the aforementioned indicator 1[15]. The fact that indicator 1 is more correlated with the factor associated with flexibility is not strange; to be able to control external conditions can be thought of as both comfortable and flexible, and in the setting imposed by the current factor model the latter is the most important.

The factors are also predicted, following the procedure described chapter 3.2.1. Estimated factors can be found in table A.4. Next, OLS regressions are conducted with the estimated factors as endogenous variables and observable char-

[13]The large difference between factors retained in that article and in this thesis is most likely due to a different estimation method for the factor loadings.

[14]This indicator loads highest on factor 3.

[15]Indicator 4 has the target dimension comfort and reads *how important is it for you to avoid changing the mode of transport?* This also have a factor loading above 0.3 for factor 3; however, it loads higher on factor 1, and should therefore be associated with factor 1 despite the high factor 3 correlation.

Table 3.3: EFA factor loadings and uniquenesses, small loadings are not displayed.

$N = 502$ Target dimensions	Indicators	Factor loadings			Specificity
		1	2	3	
Comfort	1			0.31	0.78
	2	0.66			0.52
	3	0.57			0.67
	4	0.42		0.31	0.73
Reliability	5	0.74			0.45
	6	0.55			0.66
	7	0.54			0.67
Flexibility	8			0.44	0.78
	9			0.60	0.60
	10			0.52	0.66
	11			0.69	0.52
Local environmental consciousness	12		0.53		0.71
	13		0.30		0.91
	14		0.17		0.95
	15		0.40		0.82
Safety	16		0.30		0.90
	17		0.44		0.81
	18		0.49		0.73
	19		0.34		0.87
Global environmental consciousness	20		0.16		0.96
	21		0.30		0.89
	22		0.52		0.69
	23		0.36		0.87

Note: To increase readability, only factor loadings with absolute value above 0.3 are displayed except for indicator 14 and 20. These indicators did not load above 0.3 on any of the factors, and therefore the highest loading is displayed instead. The indicator numbers refer to the numbers from the first row of table 2.2.

Table 3.4: Regression with EFA factors as endogenous variables.

$N = 493$	Factor 1 "Convenience"		Factor 2 "Carefulness"		Factor 3 "Flexibility"	
Variables	β	S.E.	β	S.E.	β	S.E.
d_female	0.22**	(0.09)	0.41***	(0.08)	-0.14*	(0.08)
d_child	-0.45***	(0.11)	-0.13	(0.10)	0.36***	(0.11)
age	-0.13***	(0.03)	0.23***	(0.03)	0.06**	(0.03)
income	0.06***	(0.02)	-0.05***	(0.02)	0.04**	(0.02)
_cons	0.31**	(0.15)	-0.93***	(0.13)	-0.50***	(0.14)
F(8, 484)	10.00		25.49		9.44	
R^2	0.08		0.17		0.07	
Adjusted R^2	0.07		0.17		0.06	

Standard errors are denoted by S.E. and written in parentheses.
* significant at 10%; ** significant at 5%; *** significant at 1%.

acteristics as explanatory variables. These are (1) a dummy for whether the respondent is a female, (2) a dummy for whether the respondent has a child, (3) age measured in decades and (4) the income of the respondent, measured in 100,000 NOK.

Variables from figure 3.4 with the prefix *d_* are dummy variables. The βs denote the marginal effects on the relevant factor by increasing the associated variables from zero to one, by ten years and by 100,000 NOK for the dummy variables, the *age* variable and the *income* variable, respectively.

It is problematic to find variables exogenous to the estimated factors; however, the only variable of the above regressors that could be endogenous to the factors is *income*. It is logical that there is some reversed causality; however, Johansson et al. (2006) argue that income can be assumed to be exogenous to latent variables, and I rely on their logic. One may also think of income as a proxy variable for a wide range of characteristics such as IQ, motivation, childhood and opportunities. These characteristics should be exogenous to the factors imposed by the model.

We see that most of the estimates are significant, however difficult to interpret apart from their sign. It seems to be more likely that young women without children have a preference towards "convenience", that old women have a preference towards "carefulness" or "political correctness" and that men with children have a preference towards "flexibility". The most surprising from the above regression is perhaps that whether you have a child or not does not seem to affect your preference for the factor related to indicators for safety (factor 2). This may be because this factor includes a wide variety of indicators which may be drawn in different directions based on the regressor values, and hence interpretation easily becomes ambiguous.

Despite significant parameters, the low R^2 values confirms the finding from Flügel (2011, p. 15), namely that attitudes seem relatively independent of socio-demographic traits. This is perhaps comforting at a human level but bad in the

context of forecasting.

3.3.3 Confirmatory factor analysis

This section describes the conducted CFA. The idea is to exploit the information that is available a priori in the estimation procedure. This is done by assuming a particular correlation structure between the indicators and the factors. The correlation structure assumed is the one displayed in table 2.2; indicators are assumed to be caused by their corresponding factor and uncorrelated to all the other factors[16]. This assumption leads to over-identification of the model system, and is more strict than it needs to be. As an example, as discussed in section 3.3.2 one might believe that indicator 1 also should be allowed to correlate with the factor representing flexibility, not only the factor representing comfort. If being able to control external conditions in fact is more important for individuals with a high preference for comfort, restricting the factor loading for comfort to zero will lead to an inconsistency.

Nonetheless, sophistication of the factor analysis by the aforementioned method of "relaxing" parameters constrained to zero comes with a cost; reduced transparency and readability, and more complex interpretations. Therefore, the CFA model is estimated in the simple form displayed in table 2.2, and sophistication is considered for the latent variables in chapter 4 instead.

Even though the model is over-identified, because of indeterminacy of the factors the effect of a factor on the first indicator associated with that factor is normalized to 1 for identification. Estimated factor loadings are displayed in table 3.5. The factor scores are then estimated and used as endogenous variables in a regression where the exogenous variables are the same as in figure 3.4. The results from these regressions are displayed in table 3.6.

The Cronbach's alpha values are $\alpha_{\eta_{comf}} = 0.56$, $\alpha_{\eta_{reli}} = 0.69$, $\alpha_{\eta_{flexi}} = 0.66$, $\alpha_{\eta_{local}} = 0.43$, $\alpha_{\eta_{safe}} = 0.50$ and $\alpha_{\eta_{global}} = 0.40$. This is a measure of the internal consistency or the reliability in the measurement of each factor based on the correlations between each of the relevant indicators, and calculated as

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \tag{3.31}$$

where k is the number of indicators used for the relevant factor and $\bar{r}$ is the average of the non-redundant correlation coefficients (i.e. the $k(k-1)/2$ terms in the top or bottom triangle) of the correlation matrix between the k indicators. A usual rule of thumb is that the Cronbach's alpha should be as high as 0.70 (Cronbach, 1951). If this rule is followed, some of the factors are borderline acceptable while most of them are unacceptable. However, as Johansson et al. (2006) points out, this is not a problem if two criteria are met; (1) if the factors from the factor analysis are only relevant as preliminary estimates of latent

[16]This means that indicator 1 is assumed to be uncorrelated to all other factors than "comfort" and indicator 5 is assumed to be uncorrelated to all other factors than "reliability", and so on.

variables in a latent variable model and (2) if low α values is a result of individual heterogeneity due to individual specific variables that are controlled for in the latent variable model (see section C.4).

Looking at table 3.6, we see that most of the parameter estimates are significant. As for EFA, the marginal effects are difficult to interpret other than through their sign. A female without children is more likely to have preferences towards *comfort* and *reliability*, while *flexibility* is associated with having children. *Environmental consciousness* (local and global) and *safety* are most likely to be preferred by older women, but the parameter reflecting age is lower in magnitude for global environmental preferences, perhaps because climate change is a relatively new concern.

As for the EFA regression, it is apparent that the R^2 values are relatively low. The CFA factors are therefore also difficult to predict based on sociodemographic traits. R^2 values are, however, somewhat higher for factors associated with behavioral indicators and this is also the case for the EFA regression. This might be because the behavioral indicators have somewhat more variation than the attitudinal indicators, but examining table A.1 it is not at all obvious that this is the case. It may also be because "political correctness" or "carefulness", which is the EFA interpretation of the behavioral factor (see section 3.3.2), is more bound by gender, age, income and/or having a child or not than preferences for comfort, flexibility and reliability are.

Table 3.5: CFA factor loadings.

$N = 502$	Factor loadings					
Indicators	η_{comf}	η_{reli}	η_{flexi}	η_{local}	η_{safe}	η_{global}
1	1.00					
	—					
2	2.36					
	(0.43)					
3	2.41					
	(0.44)					
4	1.30					
	(0.24)					
5		1.00				
		—				
6		0.83				
		(0.08)				
7		0.73				
		(0.07)				
8			1.00			
			—			
9			1.33			
			(0.19)			
10			1.42			
			(0.21)			
11			2.06			
			(0.29)			
12				1.00		
				—		
13				0.27		
				(0.05)		
14				0.23		
				(0.07)		
15				0.63		
				(0.10)		
16					1.00	
					—	
17					0.57	
					(0.13)	
18					1.25	
					(0.22)	
19					0.65	
					(0.15)	
20						1.00
						—
21						1.17
						(0.36)
22						3.39
						(0.91)
23						1.11
						(3.36)

Note: The factor loadings for indicators 1, 5, 8, 12, 16 and 21 are constrained to 1 for identification of the factors *comfort*, *reliability*, *flexibility*, *local environmental consciousness*, *safety* and *global environmental consciousness*, respectively. All cells that appear as blanks have factor loadings constrained to zero. Standard errors are reported below the factor loadings in parentheses.

Table 3.6: Regression with CFA factors as endogenous variables.

$N = 493$	η_{comf}		η_{reli}		η_{flexi}		η_{local}		η_{safe}		η_{global}	
Variables	β	S.E.	β	S.E.	β	S.E.	β	S.E.	β	S.E.	β	S.E.
d_female	0.08***	(0.03)	0.15**	(0.06)	-0.03	(0.04)	0.28***	(0.06)	0.19***	(0.04)	0.10***	(0.02)
d_child	-0.13***	(0.03)	-0.27***	(0.08)	0.17***	(0.05)	-0.08	(0.07)	-0.05	(0.05)	-0.05**	(0.02)
age	-0.03**	(0.01)	-0.09***	(0.02)	0.02***	(0.01)	0.15*	(0.02)	0.12**	(0.01)	0.04***	(0.01)
income	0.02***	(0.01)	0.05***	(0.02)	0.02	(0.01)	-0.02***	(0.01)	-0.02***	(0.01)	-0.01***	(0.00)
_cons	0.06	(0.05)	0.18*	(0.11)	-0.20***	(0.06)	-0.67***	(0.10)	-0.52***	(0.07)	-0.15***	(0.03)
F(4,488)	8.50		8.32		7.36		20.47		24.18		19.55	
R^2	0.07		0.06		0.06		0.14		0.17		0.14	
Adj. R^2	0.06		0.06		0.05		0.14		0.16		0.13	

Note: The factor name abbreviations at the top row denote *comfort*, *reliability*, *flexibility*, *local environmental consciousness*, *safety* and *global environmental consciousness*, respectively.
β denotes coefficients. Standard errors are denoted by S.E. and written in parentheses.
* significant at 10%; ** significant at 5%; *** significant at 1%.

3.4 Preliminary findings

As a basis for the latent variables in an integrated latent variable and choice model, either EFA or CFA factors can be used. The advantage with EFA is that it is more efficient; the factors explain a larger share of the total covariation. The advantage with CFA is that it incorporates constraints; if the underlying theoretical foundation or the a priori assumptions for the CFA factors are strong enough, CFA should be used. But there is always a trade-off between the amount of covariation explained and the benefits of the constraints imposed.

Section 3.3.2 shows the basis for the suggested EFA factors. If EFA factors are to be used the recommended approach is to choose a threshold value, for instance 0.3 as in table 3.3, and constrain all parameters below this threshold in absolute value to zero. This should be done for identification purposes (see section C.4). One should then check the estimated factor loadings from the integrated latent variable and choice model to see if the factors still are interpretable. If not, different threshold values should be experimented with.

If CFA factors are to be used, however, some alterations in the target dimensions that are shown in table 2.2 should be done, based on the joint findings from section 2.2 regarding the correlation structure of the indicator variables, findings from section 3.3.2 and 3.3.3 as well as the EFA conducted in Flügel (2011). Some changes from the structure in table 2.2 are proposed; in a latent variable model one should:

- Let all the factors be affected by age, gender, income and whether the respondent has a child or not;
- Allow for a relationship between the latent variables for *reliability* and *comfort*;
- Let indicator 1 be affected by *flexibility* as well as *comfort*;
- Exclude indicator 10 completely because of endogeneity problems; and
- Let indicator 12 be affected by *global environmental consciousness* as well as *local environmental consciousness*.

4 Integrated choice and latent variable model

This chapter describes a method to consistently integrate a choice model with a latent variable model, and how to estimate the model system simultaneously and fully efficiently by means of the full information maximum likelihood. Section 4.1 describes the theory behind the model framework. Section 4.2 utilizes this framework on the dataset described in chapter 2, while at the same time taking into account the preliminary conclusions regarding how to form latent variables from the indicators described in section 3.4.

4.1 Theoretical framework

This section covers the inclusion of latent variables in choice models. In other words, how to integrate a latent variable model and a choice model in such a way that the latent variables may affect the outcome of the choice. Readers not familiar with discrete choice models or latent variable models are referred to the theoretical annex. Section C.4 in that appendix covers latent variable models. Section C.5 covers discrete choice models in general and binary discrete choice models in particular. This section covers how to estimate choice models with latent variables, and is therefore the final piece of theory needed for establishing a consistent methodological framework for the high speed rail data

Taking the decision-making process (see section 3.1.1) explicitly into account when estimating choice behavior has long been deemed necessary by behavioral econometricians. The most intuitive way of doing this — running a preliminary factor, MIMIC or latent variable analysis, estimate latent variables and include them as exogenous regressors in choice models — is problematic since the latent variables will introduce measurement errors. Including them in a regression without taking measurement errors into account will result in inconsistent estimators.

However, over the last years methods for consistent incorporation of latent variables in choice models have been developed. See for instance McFadden (1986); Morikawa (1989); Ben-Akiva et al. (1999); McFadden (2000); Walker (2001); Ben-Akiva et al. (2002); Ashok et al. (2002); Johansson et al. (2006); Atasoy et al. (2010) for examples of this. In this literature, three main proce-

dures are discussed. Ben-Akiva et al. (2002) gives an overview of these methods. It is (1) possible to first obtain the distribution of factors from a latent variable or factor model and then integrate the choice probabilities over these distributions to obtain consistent but inefficient estimates, or (2) estimate the latent variable model and the choice model simultaneously by means of maximum likelihood to obtain consistent and fully efficient estimates. Since the likelihood function is a complex multidimensional integral over the distribution of latent variables that has to be solved numerically, as the number of latent variables increases the integration procedure becomes infeasible (Ben-Akiva et al., 2002, p. 12). In these cases it is possible to (3) employ simulation methods in which random draws from the estimated distribution of latent variables are collected and used to obtain an unbiased estimator. When the number of latent variables becomes large, this latter method is preferred. However, because of time constraints I focus on method (2) in the remainder of this thesis[1].

The rest of this section will present a general methodology and the main ideas behind the aforementioned literature. Since Ben-Akiva et al. (2002) contains an extensive review, the section will be based on their article unless otherwise is stated.

4.1.1 Model specification

The equation system consists of two models; a choice model and a latent variable model. Each of these models have both measurement and indicator equations. The equation system written in general form is

$$\boldsymbol{\eta} = \mathbf{h}(\mathbf{x}; \boldsymbol{\Gamma}) + \boldsymbol{\zeta} \qquad \text{and} \qquad \boldsymbol{\zeta} \sim D(\mathbf{0}, \boldsymbol{\Psi}) \tag{4.1}$$

$$\mathbf{u} = \mathbf{v}(\mathbf{x}, \boldsymbol{\eta}; \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \qquad \text{and} \qquad \boldsymbol{\varepsilon} \sim D(\mathbf{0}, \boldsymbol{\Xi}) \tag{4.2}$$

$$\mathbf{y} = \mathbf{g}(\mathbf{x}, \boldsymbol{\eta}; \boldsymbol{\Lambda}) + \boldsymbol{\xi} \qquad \text{and} \qquad \boldsymbol{\xi} \sim D(\mathbf{0}, \boldsymbol{\Theta}) \tag{4.3}$$

$$d_j = \begin{cases} 1 & \text{if} \quad u_j \geq u_s; j \neq s, \forall j,\ s \in J \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

where the first and third equation constitute the latent variable model and the second and fourth equation constitute the choice model for one individual, individual i. The individual subscript is dropped for notational convenience. The two first equations are structural equations, while the two last equations are measurement equations. We also see that equations 4.1 and 4.3 constitute the latent variable part of the model (see the top part of figure 1.1) that is described in appendix C.4 while equations 4.2 and 4.4 constitute the choice part of the model (see the left part of figure 1.1) that is described in appendix C.5.

$\boldsymbol{\eta}$ is the vector of m latent variables, $\mathbf{u}$ is the vector of utilities for the J alternatives contained in individual i's choice set, $\mathbf{y}$ is the vector of p observable indicators and the J number of d dummies indicate the utility maximizing choice

[1] If other researchers want to try the procedure described in this chapter and simultaneous maximum likelihood proves to be infeasible, see appendix C.6 for a complete wakthrough of method (1)

for individual i. These are collected in the vector $\mathbf{d}$. $\mathbf{x}$ is a vector of k observable, exogenous attributes[2]. $\boldsymbol{\Gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ are matrices of unknown parameters, $\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\xi}$ are vectors of error terms and $\boldsymbol{\Psi}$, $\boldsymbol{\Xi}$ and $\boldsymbol{\Theta}$ are covariance matrices. Typically, these covariance matrices contain numerous restrictions and normalizations for simplification and identification purposes. D denotes unspecified distributions (that need to be specified before estimation is possible) and $\mathbf{h}$, $\mathbf{v}$ and $\mathbf{g}$ are unspecified functional forms. In the application part of this chapter as well as in appendix C.4 and C.5 they are specified to be linear functions.

Equation 4.1 gives the distribution of the latent variables given the observed variables, $f_1(\boldsymbol{\eta}|\mathbf{x};\boldsymbol{\Gamma},\boldsymbol{\Psi})$. Equation 4.2 gives the distribution of utilities given latent and observed variables, $f_2(\mathbf{u}|\mathbf{x},\boldsymbol{\eta};\boldsymbol{\beta},\boldsymbol{\Xi})$. Equation 4.1 gives the distribution of indicators conditional on the distribution of the latent and the observed variables, $f_3(\mathbf{y}|\mathbf{x},\boldsymbol{\eta};\boldsymbol{\Lambda},\boldsymbol{\Theta})$.

From equations 4.2 and 4.4 and an assumption about the distribution of $\boldsymbol{\varepsilon}$ it is also straight forward to derive the J choice probabilities conditional on $\mathbf{x}$ and $\boldsymbol{\eta}$. This may for instance be done with a probit model if the error terms are independent and identically normally distributed (as done in appendix C.6) or a logit model if the error terms are independent and identically Gumbel distributed (as done in the application part of this chapter). In this section the choice model is specified with J alternatives, while in the rest of the thesis $J = 2\ \forall i$ so that binary models will be applied. Regardless of the model utilized and the assumed distribution, these probabilities are denoted $P(\mathbf{d}|\mathbf{x},\boldsymbol{\eta};\boldsymbol{\beta},\boldsymbol{\Xi})$.

4.1.2 Likelihood function

As stated earlier, the choice model may have any form and the corresponding likelihood function will be used as base for this exposition. As a starting point, the likelihood for choosing d_j for individual i when latent variables are ignored can be deduced from equations 4.2 and 4.4 and written as

$$P(d_j = 1|\mathbf{x}_j;\boldsymbol{\beta},\boldsymbol{\Xi}) = P(u_j \geq u_s; j \neq s, \forall s \in J) \tag{4.5}$$

Assuming independent error components $(\boldsymbol{\zeta},\boldsymbol{\varepsilon})$, latent variables may be included in this setup. The likelihood function is then the choice model integrated over the distribution of latent variables from equation 4.1

$$P(d_j = 1|\mathbf{x}_j;\boldsymbol{\beta},\boldsymbol{\Gamma},\boldsymbol{\Xi},\boldsymbol{\Psi}) = \int_{\boldsymbol{\eta}} P(d_j = 1|\mathbf{x}_j,\boldsymbol{\eta};\boldsymbol{\beta},\boldsymbol{\Xi}) f_1(\boldsymbol{\eta}|\mathbf{x};\boldsymbol{\Gamma},\boldsymbol{\Psi})\,\mathrm{d}\boldsymbol{\eta} \tag{4.6}$$

Indicators may be introduced if the error components ($\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{\xi}$) are assumed independent. The joint probability of the observable variables $\mathbf{y}$ and d_j conditional

[2]To simplify the exposition, only one vector of observable, exogenous attributes is included, containing both individual specific and alternative specific variables. Some variables in $\mathbf{x}$ will only affect one or two, but not all three, of the endogenous variables. This is implemented through restricting the corresponding entities of the coefficient matrices to be zero. If one is using the latent variable model framework from section C.4, $\mathbf{x}$ would be a vector containing $\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_2$.

on $\mathbf{x}$ can then be written as

$$f_4(d_j, \mathbf{y}|\mathbf{x}_j; \boldsymbol{\beta}, \Lambda, \boldsymbol{\Gamma}, \boldsymbol{\Xi}, \boldsymbol{\Psi}, \boldsymbol{\Theta}) = \int_{\boldsymbol{\eta}} P(d_j = 1|\mathbf{x}_j, \boldsymbol{\eta}; \boldsymbol{\beta}, \boldsymbol{\Xi}) f_3(\mathbf{y}|\mathbf{x}, \boldsymbol{\eta}; \boldsymbol{\Lambda}, \boldsymbol{\Theta}) f_1(\boldsymbol{\eta}|\mathbf{x}; \boldsymbol{\Gamma}, \boldsymbol{\Psi}) \, \mathrm{d}\boldsymbol{\eta} \tag{4.7}$$

The first term of the integrand corresponds to the choice model, the second term corresponds to the measurement equation and the third term corresponds to the structural equation from the latent variable model. Since the latent variables are only known to their distribution estimated from the latent variable model, the joint probability of d_j and $\mathbf{y}$ have to be integrated over the m dimensional vector $\boldsymbol{\eta}$.

Any distribution may be assumed for the disturbances in the choice model part of the likelihood function. In section 4.2 they are assumed to be independently and identically distributed and drawn from a standard Gumbel distribution. In this case the choice model part follows a standard logit model[3]:

$$\begin{aligned} P(d_j = 1|\mathbf{x}_j, \boldsymbol{\eta}; \boldsymbol{\beta}) &= P(u_j \geq u_s, \, \forall s \in J) \\ &= P(v_j + \varepsilon_j \geq v_s + \varepsilon_s, \, \forall s \in J) \\ &= P(\varepsilon_s - \varepsilon_j \leq v_j - v_s, \, \forall s \in J) \\ &= \frac{e^{v_j}}{\sum_{s \in J} e^{v_s}} \end{aligned} \tag{4.8}$$

It is conventional to assume normally and independently distributed orthogonal latent variables and normally and independently distributed indicators. This may be written as $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}\,\text{diagonal})$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}\,\text{diagonal})$, and results in the following densities for $\boldsymbol{\eta}$ and $\mathbf{y}$:

$$f_1(\boldsymbol{\eta}|\mathbf{x}; \boldsymbol{\Gamma}, \boldsymbol{\sigma}_{\boldsymbol{\zeta}}) = \prod_{l=1}^{m} \frac{1}{\sigma_{\zeta_l}} \phi\Big(\frac{\eta_l - h(\mathbf{x}; \Gamma_l)}{\sigma_{\zeta_l}}\Big) \tag{4.9}$$

$$f_3(\mathbf{y}|\mathbf{x}, \boldsymbol{\eta}; \boldsymbol{\Lambda}, \boldsymbol{\sigma}_{\boldsymbol{\xi}}) = \prod_{r=1}^{p} \frac{1}{\sigma_{\xi_r}} \phi\Big(\frac{y_r - g(\mathbf{x}, \boldsymbol{\eta}; \Lambda_r)}{\sigma_{\xi_r}}\Big) \tag{4.10}$$

where σ_{ζ_l} and σ_{ξ_r} are the standard deviations of ζ_l and ξ_r respectively, collected in the vectors $\boldsymbol{\sigma}_{\boldsymbol{\zeta}} = \text{diag}(\boldsymbol{\Psi})$ and $\boldsymbol{\sigma}_{\boldsymbol{\xi}} = \text{diag}(\boldsymbol{\Theta})$ and ϕ denotes the standard normal density function.

4.1.3 Simultaneous maximum likelihood estimation

The most efficient form of estimation is simultaneous maximum likelihood estimation. In this case, numeric integration is used to maximize the logarithm of

[3]In the case where $J = 2$, the expression in equation 4.8 will be equal to the expression in equation C.17, the equation for the binary logit.

the sample log likelihood function

$$\max_{\boldsymbol{\beta},\Lambda,\boldsymbol{\Gamma},\boldsymbol{\Xi},\boldsymbol{\Psi},\boldsymbol{\Theta}} \sum_{i=1}^{N} \ell(\boldsymbol{\beta},\Lambda,\boldsymbol{\Gamma},\boldsymbol{\Xi},\boldsymbol{\Psi},\boldsymbol{\Theta};\mathbf{y}_i,\mathbf{x}_i,\mathbf{d}_i)$$
$$= \max_{\boldsymbol{\beta},\Lambda,\boldsymbol{\Gamma},\boldsymbol{\Xi},\boldsymbol{\Psi},\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{j=1}^{J_i} d_{ij} \ln f_4(d_{ij},\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\beta},\Lambda,\boldsymbol{\Gamma},\boldsymbol{\Xi},\boldsymbol{\Psi},\boldsymbol{\Theta}) \tag{4.11}$$

where N denotes the number of individuals in the sample. Assuming that $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Psi}\,\text{diagonal})$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Theta}\,\text{diagonal})$ as in equations 4.9 and 4.10, equation 4.11 can be written as

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{N}\sum_{j=1}^{J} d_{ij} \ln\left(\int_{\boldsymbol{\eta}} P(d_{ij}=1|\mathbf{x}_i,\boldsymbol{\eta}_i;\boldsymbol{\beta},\boldsymbol{\Xi}) \prod_{r=1}^{p}\left[\frac{1}{\sigma_{\xi_r}}\phi\Big(\frac{y_{ir}-g(\mathbf{x}_i,\boldsymbol{\eta}_i;\Lambda_r)}{\sigma_{\xi_r}}\Big)\right] \right.$$
$$\left. \prod_{l=1}^{m}\left[\frac{1}{\sigma_{\zeta_l}}\phi\Big(\frac{\eta_{il}-h(\mathbf{x}_i;\Gamma_l)}{\sigma_{\zeta_l}}\Big)\right] \mathrm{d}\boldsymbol{\eta}\right) \tag{4.12}$$

where $\boldsymbol{\theta}$ is a vector of the parameters $(\boldsymbol{\beta},\Lambda,\boldsymbol{\Gamma},\boldsymbol{\Xi},\boldsymbol{\sigma_\xi},\boldsymbol{\sigma_\zeta})$. The first term can be any standard discrete choice likelihood function, depending on the distributional assumption for the error term $\boldsymbol{\varepsilon}$ (that is, the assumptions on $\boldsymbol{\Xi}$), see section C.5.

4.2 Application

In this section I will utilize the aforementioned framework and the preliminary results from chapter 3 regarding how indicators should be used in the latent variable part of the model. More specifically, I will specify the equation system 4.1–4.4 completely and estimate the system simultaneously by means of the log likelihood function written in general form in equation 4.12. To save computation time, preliminary analyses of both the choice model and the latent variable model are done in Stata 12 (StataCorp, 2011). Preliminary estimations of the latent variable models are done with the "sem" command, new in Stata 12. The whole model is then estimated simultaneously in Biogeme (Bierlaire, 2003). To estimate latent variable models in Biogeme, the new version that runs through Python must be used (Bierlaire and Fetiarison, 2009); this version allows for a more flexible specification of the likelihood function.

4.2.1 Simplifications done in the model specification

In this subsection I will describe the simplifications I was forced to make when specifying an estimable model. These simplifications are described in the bullet points below. For each of the sections below I try to explain why this simplification had to be made, and also justify the related choices I had to take during the process.

Binary choice models

The estimated models only contain choices between air and HSR for business travelers. Because of time restrictions and limited processor capacity I chose to estimate a binary choice model, and in my opinion, the most interesting market segment for HSR is business travelers that usually fly. There are three main reasons for this: (1) this is the largest potential market segment for HSR, (2) I believe this is the segment for which the highest positive socio-economic externalities can be achieved (both because business travelers' value of time (VoT) is generally considered to be higher, and because of potential positive externalities for various industries by linking the major cities in Norway closer together), and, finally, (3) because of the potential positive effects on the environment in terms of reduced carbon emissions by reducing the market share for air.

Only one latent variable

Due to numerical issues, I was not able to estimate the model based on simulation of the distributions of latent variables within the time frame. Hence, the model is estimated by means of numerical integration of the likelihood function, as is explained in section 4.1. The problem with this is that the dimension of the integral has a large effect on estimation time. Again, because of time restrictions and limited processor capacity I was only able to estimate models with one latent variable at the time.

Latent variables are based on CFA factors

There are two main approaches I have used for generating personality traits in the form of latent variables based on indicators; EFA and CFA. In my CFA analysis I nested different indicators together to form latent variables, while in the EFA analysis all latent variables are based on all indicators. Because of this, and because factors are assumed to be orthogonal, the results of my EFA are dependent on the number of factors. Hence, estimating the model with one latent variable based on EFA would alter all the factor loadings ($\mathbf{\Lambda}$) and give the latent variable a different interpretation.

I could have estimated a latent variable based on EFA by only including the indicators with a loading above a certain threshold value, for instance as in table 3.3. However, I believe environmental effects are important for the choice between air and HSR. This factor from table 3.3 is problematic since in includes questions regarding safety and global and local environmental concerns. There are two reasons for why this is problematic: (1) while safety and global environmental concerns should draw in the direction of HSR, local environmental concern should draw in the direction of air. This makes the sign of the effect ambiguous; and (2) while global environmental concern often is associated with the younger generation, concerns for safety and and the local environment are often associated with the older generation. This is a source of heterogeneity when estimating the latent variable based on observable characteristics. Since

I had limited time to experiment with such latent variables, I chose to base the latent variables on the "smaller" and more straightforward CFA factors.

Comfort and "GEC" as latent variables

I am estimating three models; one baseline model without latent variables, one model with *comfort* as a latent variable and one model with *global environmental consciousness* (GEC) as a latent variable. I will hereby justify my choice of latent variables. This has to be done in context of the market segment I am considering; business travelers that have the choice between air and HSR. *Reliability* and *flexibility* can be considered to be roughly the same for air and HSR, so these will not be discussed further.

Even if most people perceive flying to be more unsafe than other modes of transport, flying is one of the safest ways of traveling in terms of accident rates. It is also difficult to know a priori how people perceive the safety level on HSR. On one hand, it will be a new mode with the latest technology; on the other hand, it is difficult to know beforehand how things work out at the first trips. Risk averse individuals may therefore be inclined to wait an amount of time before trusting HSR to be safe. Because of this, analyzing the effect of safety is not straight forward.

It would be interesting to see the effect of *local environmental consciousness* (LEC), which is expected to draw in the direction of air. However, there are reasons to believe that LEC and GEC are highly collinear variables. This would be unproblematic if both latent variables could be estimated simultaneously; however, estimating one latent variable at the time there is a high chance that the effect of the other variable would be included unintentionally. Being mostly urban citizens I believe GEC is more important than LEC for business travelers[4]. I therefore think that the collinearity problem is gravest for the LEC variable, so that it also would reflect global concerns. That would make the sign of the effect ambiguous. For this reason I will only estimate a model with the variable GEC, for which I believe collinearity has a lesser effect[5].

Comfort is assumed to be important for two reasons: (1) for the choice between air and HSR, air would involve a lot of waiting at airports and different procedures as security checks which are inconvenient for the traveler. There are therefore reasons to believe that even with the same access and egress time,

[4]This is by no means scientifically based, only an intuitive approach. One could for instance argue that business travelers are on average older and therefore more reluctant to care about the global environment. One could also argue that they should value global environment less since being businessmen they must appreciate industry development. On the other hand, a lot of industries has recently discovered the marketing value of being green. If one is interested in finding out if GEC og LEC is most important for business men, this should obviously be investigated further.

[5]It is not certain whether the stated choice of mode reflects the preferences of the individual or the preferences of the company; however, because of the way the questions are formulated and the fact that individuals answer the questionnaires in their own name, there are reasons to believe the stated choices are done at an individual level. Even if the choices reflect the policy of the company, an individual's environmental consciousness is likely to be correlated with the environmental consciousness of the company for which she is working.

comfort would draw in the direction of HSR. And (2), business travelers are the passenger segment that is most likely to value comfort, since (a) the cost of comfort is often fully covered by the company and (b) business travelers often travel more then others, and therefore the time aboard transport modes is relatively more important. Hence, comfort is assumed to affect business travelers more than other people.

4.2.2 Model specification

This section describes the complete specifications of three models for which the motivation can be found in the previous section. The first model is a binary choice between HSR and air. The second model incorporates the latent variable "comfort". The third model incorporates the latent variable "global environmental consciousness". The binary model is a logit model, and the reason for this is to be able to calculate marginal effects as explained in footnote 7. The only deviation from the aforementioned framework is that the constant terms below are written explicitly as αs instead of being part of other parameter vectors.

Model 1: The baseline model

The below model is a binary logit model between HSR and air. The binary logit framework is achieved by assuming that the error terms ε_{AIR} and ε_{HSR} are Gumbel distributed. The CDF of the Gumbel distribution is $G(x) = e^{-e^{-x}}$ and the variance is $\pi^2/6$. The difference between two independent, Gumbel distributed variables is logistically distributed with mean zero and variance $\pi^2/3$. Hence, by defining a new function that is the difference between the two utility functions ($u_{AIR} - u_{HSR}$), the error term for the new function ($\varepsilon_{AIR} - \varepsilon_{HSR}$) will then follow a logistic distribution. The chosen model specification is

$$u_{AIR} = \boldsymbol{\beta}'_{AIR}\mathbf{x}_{AIR} + \varepsilon_{AIR} \tag{4.13}$$

$$u_{HSR} = \alpha_{HSR} + \boldsymbol{\beta}'_{HSR}\mathbf{x}_{HSR} + \varepsilon_{HSR} \tag{4.14}$$

where $\boldsymbol{\beta}_{AIR}$ is a 4 dimensional row vector, $\boldsymbol{\beta}_{HSR}$ is an 8 dimensional row vector, $\mathbf{x}_{AIR}$ = (tidomb_ref, totkost_ref, (tidtil_ref + tidfra_ref), avg_ref)' and $\mathbf{x}_{HSR}$ = (tidomb_hht, totkost_hht, (tidtil_hht + tidfra_hht), avg_hht, tunnel_hht, age, income, d_female)'. These variables are explained in table 2.4; however, they differ from the variables in the table with respect to scaling. The age variable is divided by 10, all time and cost variables are divided by 100, and the income variable is divided by 100,000. This is done to ease estimations[6], and must be taken into account when interpreting the sizes of the estimated effects. The vectors $\mathbf{x}_{AIR}$ and $\mathbf{x}_{HSR}$ will be the same in all models.

Looking at the $\mathbf{x}$s, it becomes apparent that I have chosen to include all variables available in the utility function except for d_child. This is because

[6]The optimization algorithms of Biogeme work better when estimated coefficients are as close to 1.0 in absolute value as possible.

children are rarely brought and should therefore not affect business trips. I have also chosen to include income even though business trips rarely are paid for by the respondents themselves. This can be justified in two ways: (1) it is likely that respondents answered the questionnaire on a personal basis and not on behalf of the company; in this case income should be included, and (2) if respondents answered on behalf of the company, their income may reflect what their company can afford to use on their travels.

In addition, the cost coefficients are assumed to be generic; that is, the restriction $\beta_{AIR2} = \beta_{HSR2}$ is imposed, where these β_2s are the coefficients for totkost_ref and totkost_hht, respectively. This is done because the utility cost of paying a certain amount of NOK is assumed to be the same, whether it is used on HSR or air.

Looking at the vectors of observable variables, one sees that access and egress time also are assumed to affect the utility equally much. This is empirically established based on preliminary analyses, and also seems to be intuitively correct since most business travelers have to travel both ways. These assumptions on the $\boldsymbol{\beta}$ vectors are the same in all three models.

Model 2: Incorporating comfort

This model incorporates the latent variable *comfort* by the aforementioned framework. ε_{AIR} and ε_{HSR} are still assumed to be Gumbel distributed, $\zeta_{comf} \sim N(0, \sigma^2_{\zeta_{comf}})$ and $\xi_i \sim N(0, \sigma^2_{\xi_i})$, $\forall i \in [1, 4]$. All error terms ε_{AIR}, ε_{HSR}, ζ_{comf}, ξ_1, ξ_2, ξ_3 and ξ_4 are assumed to be independent. The model system is

$$u_{AIR} = \boldsymbol{\beta}'_{AIR}\mathbf{x}_{AIR} + \varepsilon_{AIR} \tag{4.15}$$

$$u_{HSR} = \alpha_{HSR} + \boldsymbol{\beta}'_{HSR}\mathbf{x}_{HSR} + \beta_{comf}\eta_{comf} + \varepsilon_{HSR} \tag{4.16}$$

$$\eta_{comf} = \alpha_{comf} + \boldsymbol{\Gamma}'\mathbf{x} + \zeta_{comf} \tag{4.17}$$

$$y_1 = \eta_{comf} + \xi_1 \tag{4.18}$$

$$y_2 = \alpha_2 + \lambda_2\eta_{comf} + \xi_2 \tag{4.19}$$

$$y_3 = \alpha_3 + \lambda_3\eta_{comf} + \xi_3 \tag{4.20}$$

$$y_4 = \alpha_4 + \lambda_4\eta_{comf} + \xi_4 \tag{4.21}$$

where $\boldsymbol{\beta}_{AIR}$, $\boldsymbol{\beta}_{HSR}$, $\mathbf{x}_{AIR}$ and $\mathbf{x}_{HSR}$ are defined as before, $\boldsymbol{\Gamma}$ is a 3 dimensional row vector and $\mathbf{x}$ = (age, income, d_female)'. Note that it should be desirable to include number of children in $\mathbf{x}$, since this vector is supposed to reflect the underlying personality and not preferences directly related to the trip. However, the variable d_child only reports whether the respondent brought a child on the reference trip, and only looking at business trips the variable is likely to have little variation. The comfort variable is normalized by use of indicator 1.

Model 3: Incorporating global environmental consciousness

This model is similar to the one above, but instead of comfort it incorporates *global environmental consciousness* in the utility function. The variances are

defined the same way; ε_{AIR} and ε_{HSR} are still assumed to be Gumbel distributed, $\zeta_{global} \sim N(0, \sigma^2_{\zeta_{global}})$ and $\xi_i \sim N(0, \sigma^2_{\xi_i})$, $\forall i \in [12, 20, 21, 22, 23]$. All error terms ε_{AIR}, ε_{HSR}, ζ_{comf}, ξ_{12}, ξ_{20}, ξ_{21}, ξ_{22} and ξ_{23} are assumed to be independent. In addition to the four GEC indicators from table 2.2, indicator 12 is included based on section 3.4. This gives the equation system

$$u_{AIR} = \boldsymbol{\beta}'_{AIR}\mathbf{x}_{AIR} + \varepsilon_{AIR} \tag{4.22}$$
$$u_{HSR} = \alpha_{HSR} + \boldsymbol{\beta}'_{HSR}\mathbf{x}_{HSR} + \beta_{global}\eta_{global} + \varepsilon_{HSR} \tag{4.23}$$
$$\eta_{global} = \alpha_{global} + \boldsymbol{\Gamma}'\mathbf{x} + \zeta_{global} \tag{4.24}$$
$$y_{12} = \eta_{comf} + \xi_{12} \tag{4.25}$$
$$y_{20} = \alpha_{20} + \lambda_{20}\eta_{global} + \xi_{20} \tag{4.26}$$
$$y_{21} = \alpha_{21} + \lambda_{21}\eta_{global} + \xi_{21} \tag{4.27}$$
$$y_{22} = \alpha_{22} + \lambda_{22}\eta_{global} + \xi_{22} \tag{4.28}$$
$$y_{23} = \alpha_{23} + \lambda_{23}\eta_{global} + \xi_{23} \tag{4.29}$$

where $\boldsymbol{\beta}_{AIR}$, $\boldsymbol{\beta}_{HSR}$, $\boldsymbol{\Gamma}$, $\mathbf{x}_{AIR}$, $\mathbf{x}_{HSR}$ and $\mathbf{x}$ are defined as before. The GEC variable "global" is normalized by use of indicator 12.

4.2.3 Estimation process and related weaknesses

Model 1 is estimated as equation 4.8, while models 2 and 3 are estimated based on the log likelihood function 4.12, where $P(d_{ij} = 1|\mathbf{x}_{AIR}, \mathbf{x}_{HSR}, \eta)$ is a logit function as defined in equation 4.8, the functions $\mathbf{g}$ and $\mathbf{h}$ are defined as above and the vector of parameters which the likelihood function is maximized with respect to is $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \boldsymbol{\sigma_\xi}, \boldsymbol{\sigma_\zeta})$[7]. The parameter estimates $\hat{\boldsymbol{\theta}}_{ML}$ are obtained and estimation results are displayed in table 4.1 for all parameters except α_i, σ_{ξ_i}, $\forall i \in [1, 2, 3, 4, 12, 20, 21, 22, 23]$ in models 2 and 3. These parameters are not thought of as being important for the interpretation. In the remainder of this section two weaknesses of the estimation procedure which gravely undermines the relevance of the estimates will be described.

The panel structure

Each respondent has conducted 14 different stated choices. I will call this a "panel structure" in the remainder of this thesis even though there is no time dimension in the dataset. This panel structure should have two implications: (1) the information should be taken into account when estimating the full model, and (2) latent variables should only be predicted once for each respondent, not one for each observation. Due to time constraints, neither of this was done. This will lead to perfect collinearity between the choices for each indicator for each respondent. It also means that even if we know that choices taken by the

[7] The vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have not been defined previously in this chapter; $\boldsymbol{\alpha}$ it is a collection of all the constant terms in the models (all αs) and similarly, $\boldsymbol{\beta}$ is a collection of all coefficients from the utility functions (all βs).

same individual should be more equal, this information cannot be taken into account in the estimation process.

I estimated model 2 in an alternative way, where individual specific draws were generated from the estimated distribution of the variable comfort and where each utility function included an individual specific error term with generic variance. This error term then measures the degree of variability in the choices by each individual conditional on the specified utility functions. In this model the final log likelihood increased from about -10,000 to about -5,000. However, Biogeme failed to compute the Hessian matrix and hence no standard errors for the estimated coefficients were reported. Therefore I rejected the estimates. Due to time constraints I was not able to solve this problem, and therefore all coefficients in table 4.1 are estimated as if there is one individual per stated choice. This is an indication, however, that by taking the panel structure into account one could greatly improve the explanatory power of the model.

The null log likelihood

The null log likelihood value is defined as the log of the likelihood for observing the actual choices given that all individuals chooses at random (50% chance of choosing air and 50% chance of choosing HSR) and is often used as a baseline to compare with final log likelihood values. R^2 and $adj.R^2$ values are not meaningful when it comes to discrete choice. However, two measures comparable to the measures above that are based on likelihood values to represent goodness of fit are ρ^2 and $\bar{\rho}^2$, respectively. These are by default reported by Biogeme. ρ^2 is defined as follows:

$$\rho^2 = 1 - \frac{\ell(final)}{\ell(null)} \tag{4.30}$$

where $\mathcal{L}(\cdot)$ is the log likelihood value of the model and the null log likelihood, respectively. In my thesis I want to see if the introduction of latent variables reduces individual heterogeneity. Ideally this can be checked by observing whether the ρ^2 value increases significantly from the baseline model. However, looking at the null log likelihood values from table 4.1 the problem becomes apparent; when latent variables are included the final log likelihood seems to be significantly smaller than the null log likelihood.

The reason is that the null log likelihood is automatically calculated based on 50% probability for each choice for each observation[8]. In model 1, this is comparable to the final log likelihood, which is calculated as $\sum_i^N lnP(d|\mathbf{x}, \hat{\boldsymbol{\theta}})$ where I by $\mathbf{x}$ mean all variables that are conditioned on. In models 2 and 3, however, the final log likelihood is the joint probability of observing the choice of mode as well as the choice of indicator values for all individuals, that is $\sum_i^N lnP(d, \mathbf{y}|\mathbf{x}, \hat{\boldsymbol{\theta}})$ which naturally is a smaller number and hence not comparable to the null log likelihood reported.

[8]That is the log of the probability of observing the actual outcomes $P(d) = 0.5$ multiplied together for each observation, namely $1851 \times ln(0.5)$ in models 1 and 2.

This means that I don't have any means of calculating goodness of fit statistics for the latent variable models. For this reason, these models have the ρ^2 values excluded in table 4.1[9]. Hence, it is problematic to evaluate the explanatory power of models 2 and 3 compared to model 1.

4.2.4 Estimation results

Estimated coefficients are displayed in table 4.1. The first thing that should be noted is that both *comfort* and *global environmental consciousness* are significant variables at the 1% level and have the expected signs. This is a clear indication that such personality traits play a role in the choice process and the most important result from this section. The rest of the section will go more in-dept on the estimated coefficients.

Interpreting the estimates from these models is only a partly meaningful exercise since I have not spent a great deal of time with the model specification. However, a short discussion is in place. First, I will briefly comment on the regression statistics, i.e. the bottom rows of table 4.1. The reason model 3 has fewer observations is that the indicator variables for this model has some missing values. This should also partly explain why this model has a lower final log likelihood than model 2 has. We see that the reference model has a ρ^2 of 0.17. This is a measure of goodness of fit; however, it is not strictly useful when there are no other ρ^2s to compare with. See section 4.2.3 for why the other ρ^2s are blank.

All parameters in the utility functions that are significant at 10% or less have the expected signs. The only exception is perhaps income; a negative sign on income which is a variable affecting the utility for HSR means that a higher income should reduce the demand for HSR. This is strange since table 2.5 shows that on average, HSR is the more expensive alternative[10]. However, it is in my opinion not worth reflecting much on this matter, since the model could be better specified and the estimated effect of income is negligibly small (remember also that income is measured in 100,000 NOK).

Marginal effects from logit models are easy to calculate; remember from section C.5 that the marginal effect of an increase in variable x_h is $G(\boldsymbol{\beta}'\mathbf{x})(1 - G(\boldsymbol{\beta}'\mathbf{x}))\beta_h = P(\cdot)(1 - P(\cdot))\beta_h$[11]. When calculating marginal effects, it is also important to remember the scaling of the variables from section 4.2.2.

[9] I could have excluded null log likelihood values as well since they are meaningless in models 2 and 3, but chose to include them for this point to be better illustrated.

[10] It could be many reasons for why this is the case. For instance, perhaps people with a high ethical or moral standard choose jobs based on other criteria than income, and this group is (and the companies they are working for are) more inclined to travel by HSR to save the environment.

[11] In model 1 and model 2 there are 704 stated choices for air and 1147 stated choices for HSR. In model 3 there are 532 stated choices for air and 815 stated choices for HSR. This means that for estimating the average partial effects across the whole sample, the value $P(1-P) = 0.236$ should be used for model 1 and 2, and the value $P(1-P) = 0.275$ should be used for model 3. For estimating the marginal effect for a particular group $G(\hat{\boldsymbol{\beta}}'\mathbf{x})(1 - G(\hat{\boldsymbol{\beta}}'\mathbf{x}))\hat{\beta}_h$ should be calculated for the appropriate value of each x. However, this will not be done here.

Table 4.1: Regression results.

$\hat{\boldsymbol{\beta}}$:	(1)		(2)		(3)	
Variables	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
tidomb_ref	-0.911***	0.131	-0.856***	0.131	-1.03***	0.165
tidomb_hht	-1.36***	0.142	-1.39***	0.147	-1.33***	0.162
totkost_ref	-0.119***	0.0150	-0.123***	0.0161	-0.104***	0.0159
totkost_hht	-0.119***	0.0150	-0.123***	0.0161	-0.104***	0.0159
tidtil_ref	-1.00***	0.175	-0.885***	0.178	-1.30***	0.202
tidtil_hht	-1.07***	0.157	-0.991***	0.159	-1.25***	0.188
tidfra_ref	-1.00***	0.175	-0.885***	0.178	-1.30***	0.202
tidfra_hht	-1.07***	0.157	-0.991***	0.159	-1.25***	0.188
avg_ref	-0.0140	0.0136	-0.00886	0.0140	-0.00727	0.0193
avg_hht	0.0640***	0.0135	0.0623***	0.0140	0.0456***	0.0185
tunnel_hht	-0.00122	0.00365	0.00211	0.00374	0.00574	0.00439
age	0.0399	0.501	0.210	0.536	-0.515	0.630
income	-6.22e-07*	3.44e-07	-8.63e-07**	3.57e-07	-9.35e-07**	4.22e-07
d_female	0.117	0.124	0.0726	0.131	-0.0334	0.172
_cons_hht	1.02**	0.501	-2.73***	0.809	-1.14	0.809
η_{comf}	—	—	1.17***	0.197	—	—
η_{global}	—	—	—	—	0.598***	0.154
$\hat{\boldsymbol{\Lambda}}$:	(1)		(2)		(3)	
Indicators	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
y_1	—	—	1.00	—	—	—
y_2	—	—	1.43***	0.126	—	—
y_3	—	—	1.45***	0.143	—	—
y_4	—	—	1.30***	0.109	—	—
y_{12}	—	—	—	—	1.00	—
y_{20}	—	—	—	—	0.729***	0.0880
y_{21}	—	—	—	—	0.604***	0.0781
y_{22}	—	—	—	—	1.23***	0.130
y_{23}	—	—	—	—	0.899***	0.0691
$\hat{\boldsymbol{\Gamma}}$:	(1)		(2)		(3)	
Variables	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
age	—	—	-0.0872	0.128	0.623***	0.186
income	—	—	0.0175***	0.00660	0.0656***	0.0173
d_female	—	—	0.0636**	0.0263	0.531***	0.0397
_cons	—	—	3.34***	0.0606	3.02***	0.152
σ_ζ	—	—	0.386***	0.0280	0.529***	0.0436
N :	1851		1851		1347	
Parameters:	12		28		31	
Null LL:	-1283.02		-1283.02†		-933.669†	
Final LL:	-1061.722		-10646.950		-10086.689	
ρ^2 :	0.172		—		—	
$\bar{\rho}^2$:	0.163		—		—	

Note: The columns denoted S.E. contain robust asymptotic standard errors. * significant at 10%; ** significant at 5%; *** significant at 1%.
†These values are wrong, but included for illustrational purposes; see section 4.2.3 for more information.

Looking at effects of marginal changes in latent variables first, we notice that the scale of the latent variables do not have a clear interpretation. Therefore I will calculate the effect of changes in the magnitude of one standard error. According to model 2, the effect of increasing *comfort* by one standard error is a 10.7% increase in the demand for HSR. According to model 3, the effect of increasing *global environmental consciousness* by one standard error is an 8.70% increase in the demand for HSR. That comfort seems most important for business travelers is perhaps something that complies with our intuition; however, it could also be the case that this is the result of an endogeneity bias, as explained in section 3.1.3.

Next, we turn to effects of some observable variables. We see that increasing the in-vehicle time for air by 10 minutes would increase the demand for HSR by 2.0%, and reducing the in-vehicle time for HSR by 10 minutes would increase the demand for HSR by 3.3% according to model 2. These numbers are 2.8% and 3.7% for model 3, respectively. Similarly, increasing cost of an air ticket or reducing cost of a HSR ticket by 100 NOK would increase demand for HSR by 2.9% according to both model 2 and model 3. It is also worth noticing that including latent variables does not change any marginal effects of observable variables significantly from the baseline model; however, point estimates changes slightly.

Looking at the estimated parameters in $\hat{\mathbf{\Gamma}}$ we see that being a woman and having a high income is supposed to increase respondent's preferences for both comfort and the environment. Also, perhaps surprisingly, the older the respondent is, the more he seems to care about the global environment; being 10 years older increases GEV by 1.18 standard errors. Being female increases GEC by 1.00 standard error. When it comes to comfort, being female increases the variable with 0.17 standard errors. Based on the estimates, the comfort variable seems more independent of socio-economic characteristics than the GEC variable. Comfort also has a smaller variance. This indicates that "comfort" is more stable than GEC throughout the sample of business travelers.

Comparing the estimates with the OLS regression from table 3.6, we see that the effects have the same sign for all variables, except for the effect of income on GEV which is now positive in the integrated model. This could be because we are only looking at a subset of business travelers, while table 3.6 was a regression on the whole sample. It could also be because of some collinearities between income and d_child, which is excluded from the regressions in this chapter. However, this is unlikely, since the variable d_child should be almost negligible for business travelers. The sizes of the $\hat{\mathbf{\Gamma}}$ estimates are not comparable to the sizes of the estimated coefficients from table 3.6. This is because the unit of each latent variable not necessarily is the same as in table 3.6; an indication of this is that the sizes of the estimates in $\hat{\mathbf{\Lambda}}$ in table 4.1 are completely different from the sizes of the estimates in table 3.5. Part of the reason is that in this chapter GEC is normalized with respect to indicator 12. In chapter 3 it was normalized with respect to indicator 20. It would be interesting to see the R^2 values for the equations for the latent variables and compare them with the R^2 values from table 3.6; however, they are not available.

5 Suggestions for further research

The previous chapter presented an application of the proposed method; however, the model is far too simplistically specified for the results to be relevant for e.g. policy implications for HSR in Norway. If one is interested in conducting further latent variable analyses on the case of HSR in Norway it is nonetheless possible to get more realistic results within the same model framework. This chapter contains suggested improvements to the model application from the previous chapter. Section 5.1 contains extensions of the choice model part and section 5.2 contains extensions of the latent variable model part.

5.1 Choice model extensions

This section contains choice model extensions applicable for the dataset. I will not go deep into this since different choice models for HSR in Norway already are proposed by Flügel et al. (2012). However, I will briefly describe the intuition behind the models they are proposing. The model in chapter 4 only contains stated choices between air and HSR even though the dataset also includes the modes car, bus and conventional rail. If one is interested in predicting the demand for the hypothesized mode HSR, the first extension of my model should be to include all modes of transport in the analysis. There are, however, many different types of models for which this can be achieved, which all imposes various assumptions and restrictions. I will start with the most intuitive one and end with the most sophisticated one. I describe four different models, all of which are proposed for the dataset I am considering by Flügel et al. (2012).

The *conditional logit model* (McFadden, 1973) is an extension of the binary logit model with more alternatives. In this model, the choice between two of the alternatives coincides with the binary logit model. This highlights an important property of these models; the independence of irrelevant alternatives (IIA) assumption. This assumption states that the odds ratio for choosing between two alternatives only depends on the attributes of these alternatives, and is unaffected of anything that may happen with other alternatives. In other words, this restricts all error terms to be uncorrelated. This is less problematic when the aforementioned personality traits are controlled for; even so, one might expect that there are still some characteristics not controlled for that co-varies positively between the different modes of transport. This would violate the IIA

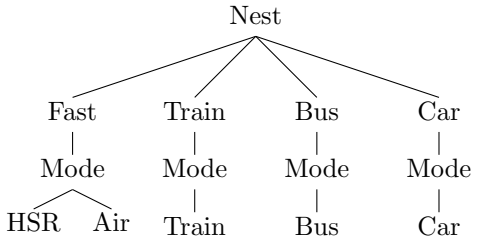

Figure 5.1: Potential nest structure for a NL model 1.

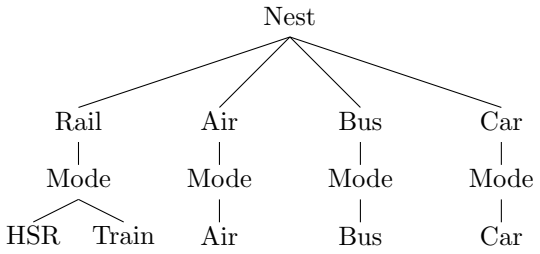

Figure 5.2: Potential nest structure for a NL model 2.

property, and the conditional logit model would thus be inconsistent.

McFadden later showed how the conditional logit model only is a special case of the family of generalized extreme value (GEV) models (McFadden, 1978). By assuming other structures for the error terms that comply with the requirements for GEV models it is possible to partly get around the IIA restriction.

The scale parameters in GEV models are parameters inversely related to the variance, and assumed to be equal to one in the conditional logit model. The *heterostscedastic logit model* (Train, 2003, section 4.5) allows for different scale parameters for different user groups or alternatives, provided that (1) one scale parameter is normalized to e.g. 1, and (2) some of the coefficients are generic for all alternatives. This means that users of different modes of transport can have different error variances. Hence, modes of transport "poorly explained" by observable variables and personality traits can be given a greater variance. Therefore it is reasonable to expect that the heteroscedastic logit model provides a better fit than the conditional logit model does.

While the heterostcedastic logit model is able to incorporate different error variances for different modes of transport, the *nested logit (NL) model* is able to nest together alternatives that are closer in the choice process (Train, 2003, section 4.2). The IIA assumption is relaxed in that it still needs to hold between nests, but not within nests. If the nested logit model is chosen, one needs to formulate a smart nesting structure so that the most equal alternatives are nested together. Atkins (2012a,b) estimated a nested logit where HSR was nested together with air (see figure 5.1). The idea was to nest the fast modes together. However, Flügel et al. (2012, p. 14, footnote 16) argues that since

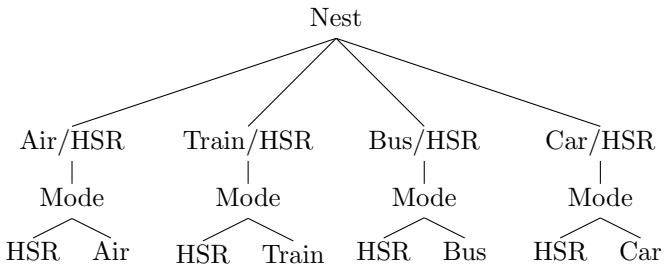

Figure 5.3: Potential nest structure for a CNL model.

travel time, travel cost and trip purpose (i.e. working trip or non-working trip) are attributes that are controlled for in both the Atkins and the TØI analysis, the conditional utility of HSR is more likely to be correlated with for instance conventional rail than air (see figure 5.2).

A more flexible model than the NL model is the *cross-nested logit (CNL) model* (Papola, 2004; Bierlaire, 2006; Abbe et al., 2007; Flügel et al., 2012). In the CNL framework the alternatives are allowed to appear in multiple nests, and hence, the nesting structure can be made more flexible. Flügel et al. (2012) propose a nesting structure as in figure 5.3 in which choice of nest is equal to the RP choice and each nest correspond to the SP choices conditional on the RP choice. In this way the CNL model captures the structure of the choice set of the individuals as it is shown in table 2.1. However, for this structure the RP choices are not allowed to be correlated.

Considering that some parameters in the CNL model do not seem to be intuitively interpretable in the behavioral context of utility maximization, some may claim that the CNL model is an unattractive choice. Train (2003, p. 98) argues that this is not necessarily a disadvantage:

> "The lack of intuition behind the properties [of GEV models] is a blessing and a curse. The disadvantage is that the researcher has little guidance on how to specify a [density function] that provides a model that meets the needs of his research. The advantage is that the purely mathematical approach allows the researcher to generate models that he might not have developed while relying only on his economic intuition."

5.2 Latent variable model extensions

Subsection 5.2.1 contains a description of the structure of a model in which more latent variables are included. The problems with this are mainly related to longer computation time, and therefore appendix C.6 describes a two-step estimation procedure so that inclusion of many latent variables may become feasible. Subsection 5.2.2 discusses the problem that arises from ordinal indicator values and proposes the ordered logit model as an example of a solution.

5.2.1 Including more latent variables

All estimations in the previous chapter only contain one attitude or personality trait variable. This was done because of time and hardware constraints; simultaneous likelihood estimation of the integrated latent variable and choice model is computationally demanding. The most obvious way of expanding the analysis is by including all the hypothesized personality traits (see e.g. table 2.2) as latent variables. This is possible within the general theoretical framework described in section 4.1, so there is no need to expand any theory here.

Figure 5.4 displays a proposed model structure where all six hypothesized personality traits are included, as well as the indicators they are supposed to affect according to section 3.4. If one wants to include causal relationships between the latent variables as well, this can be done by the method described in footnote 2 and footnote 4 from appendix C. According to section 3.4 a model extended in this manner should include a link between the personality traits *reliability* and *comfort*.

Finally, if such estimation proves to be infeasible because of hardware constraints, an alternative estimation method is described in appendix C.6. This method is not computationally demanding; it estimates personality traits first, and then includes them in the choice model in a fully consistent way. This is nonetheless a step back in terms of efficiency; when the ML estimation is not simultaneous, some of the information is not taken into account when estimating the first step. Therefore the estimation method described in section 4.1 is preferable.

5.2.2 Taking the ordinal indicator structure into account

The indicator variables described in section 2.2 are the foundation on which the whole framework for these kinds of latent variable models build. As previously described, these are questions where the respondent answers on a scale from 1 to 5, and in chapter 4 they are treated as ordinary continuous variables. The most important implicit assumption one then makes is to treat the values as meaningful numerically (i.e. that 4 is twice as high as 2). If this is not the case, the whole analysis is invalid.

Instead of treating the indicator variables as numerical values, one could treat them as ordinal values. This means that only the order of the numbers is meaningful. A class of models for analyzing this is ordered response models. Ordered response models are similar to binary choice models, but instead of one threshold value to decide whether d equals to zero or one (as in section C.5), several threshold values are used. Using the previous notation, we have the indicators $y_i,\ i \in [1, \ldots, p]$ where i denotes the indicator in question. Their value are assumed to be influenced by the m dimensional vector of latent variables, $\boldsymbol{\eta}$. This can be represented by the equations (suppressing individual specific subscripts)

$$y_i^* = \boldsymbol{\lambda}_i' \boldsymbol{\eta} + \xi_i \tag{5.1}$$

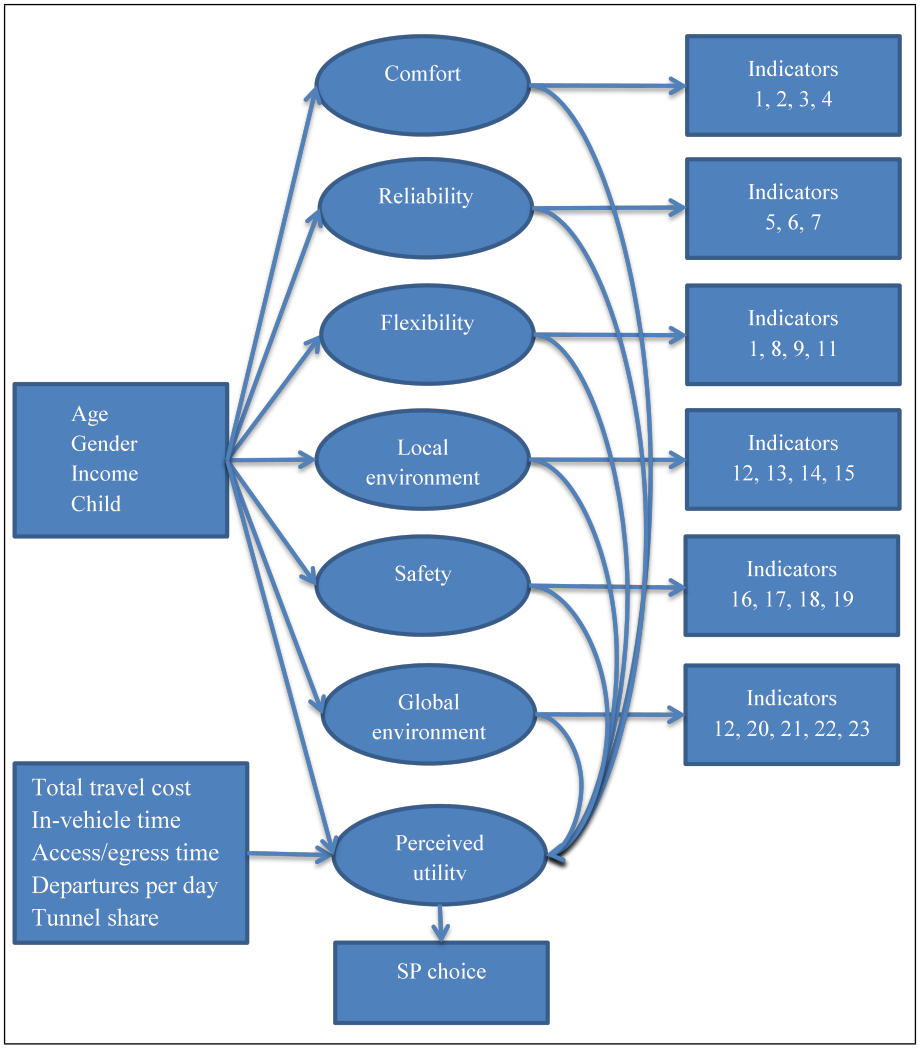

Figure 5.4: Integrated latent variable and choice model with all six personality traits included.

where y_i^* is a latent construct analogous to the utility and decided by the personality traits and an error term and $\boldsymbol{\lambda}_i$ denotes the p dimensional parameter vector for equation i. Letting $P(\xi_i \leq a|\boldsymbol{\eta}) = G(a)$ and defining the thresholds $\tau_{i1}, \ldots, \tau_{i4}$ so that the observed outcome is

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* \leq \tau_{i1} \\ 2 & \text{if} \quad \tau_{i1} < y_i^* \leq \tau_{i2} \\ 3 & \text{if} \quad \tau_{i2} < y_i^* \leq \tau_{i3} \\ 4 & \text{if} \quad \tau_{i3} < y_i^* \leq \tau_{i4} \\ 5 & \text{if} \quad \tau_{i4} < y_i^* \end{cases} \tag{5.2}$$

This gives the choice probabilities

$$\begin{aligned} P(y_i = 1|\boldsymbol{\eta}) &= G(\tau_{i1} - \boldsymbol{\lambda}_i'\boldsymbol{\eta}) \\ P(y_i = 2|\boldsymbol{\eta}) &= G(\tau_{i2} - \boldsymbol{\lambda}_i'\boldsymbol{\eta}) - G(\tau_{i1} - \boldsymbol{\lambda}_i'\boldsymbol{\eta}) \\ P(y_i = 3|\boldsymbol{\eta}) &= G(\tau_{i3} - \boldsymbol{\lambda}_i'\boldsymbol{\eta}) - G(\tau_{i2} - \boldsymbol{\lambda}_i'\boldsymbol{\eta}) \\ P(y_i = 4|\boldsymbol{\eta}) &= G(\tau_{i4} - \boldsymbol{\lambda}_i'\boldsymbol{\eta}) - G(\tau_{i3} - \boldsymbol{\lambda}_i'\boldsymbol{\eta}) \\ P(y_i = 5|\boldsymbol{\eta}) &= 1 - G(\tau_{i4} - \boldsymbol{\lambda}_i'\boldsymbol{\eta}) \end{aligned}$$

If one assumes that ξ_i is logistically distributed so that $G(\cdot)$ is the CDF of the logistic distribution (see section C.5), the model is an ordered logit model that can be estimated by means of maximum likelihood to get the estimates $\hat{\boldsymbol{\lambda}}_{iML}, \hat{\tau_{i1}}_{ML}, \ldots, \hat{\tau_{i4}}_{ML}$. Doing this for all $i \in [1, \ldots, p]$ should give the optimal threshold values. These p equations can then be included in the model system described in chapter 4 instead of the equations in 4.3 so that the whole system can be estimated simultaneously.

For completeness's sake, it should also be noted that a less ambitious approach is to use a linear, continuous function, but experimenting with different values. Substituting the values $(1, 2, 3, 4, 5)$ with the values $(1, 3, 4, 5, 7)$ would give a higher weight to those who have answered 1 or 5. In the same way, using the values $(1, 2, 4, 6, 7)$ would give less weight to those who have answered 1 or 5.

6 Conclusions

This thesis has two primary intentions: (1) to give a synthesis of the relevant theory for including personality traits as latent variables in choice models and (2) to describe a case study in which this latent variable framework is used to calculate choice probabilities. This is done for the hypothesized mode high speed rail in Norway. The considered legs are Oslo-Trondheim and Oslo-Bergen and the market segment chosen is business travelers that today use the mode air. These choice probabilities are crucial components when estimating the demand for HSR and also when predicting the reduction in the market share for air if HSR is to be built.

I predict variables for personality traits to reflect preferences for *comfort* and *global environmental consciousness*. Personality traits are thought of stable, underlying factors that affect preferences. Hence, including personality traits as latent variables in choice models should significantly reduce individual heterogeneity and at the same time benefit the behavioral interpretation of the coefficients.

I find that the latent variables are significant in the choice process, and even more importantly, they seem to be better predictors of the outcome than conventional individual-specific socio-economic variables as gender, age and income. This indicates that individual heterogeneity is reduced. Due to lack of appropriate null log likelihood values I was unfortunately not able to calculate a goodness of fit statistic that measures the explanatory power of the models within the time frame of the thesis. However, evidences from other case studies[1] are unambiguous in that inclusion of latent variables improves goodness of fit. I therefore argue that this is something that should be looked further into for HSR in Norway as well.

In addition to reducing individual heterogeneity the model framework makes it possible to understand how different individual specific characteristics affect the personality traits (the top, left part of figure 1.1). This allows for predicting different personality traits for different segments of individuals, and hence one should be able to predict the distribution of personality traits over the whole population. This is of particular interest in the context of forecasting.

A lesson learned is that it is difficult to find observable variables that are

[1] See for instance the three case studies described in Walker (2001), two case studies described in Ashok et al. (2002) as well as one case study in Morikawa (1989), one case study in Johansson et al. (2006) and one case study in Atasoy et al. (2010).

good predictors of personality traits. Hence, a recommendation is that when designing a survey, care must be taken to figure out the relevant parts of the decision making process one wants to model as latent variables and also which observable attributes that may predict these latent variables[2].

The particular contribution of this thesis can be summarized by the three following points. Firstly, this thesis synthesizes all relevant theory; both regarding factor analyses, discrete choice models, latent variable models and a consistent framework in which latent variables enter the choice model. These theory sections are by no means my own work; all theory sections are summaries of other sources and references are included in the appropriate places. However, it is to my knowledge no other sources in which all this theory is collected. In this manner my thesis provides added value for researches wanting to analyze choices in an attitudinal context since it describes the complete theoretical foundation of all the related processes.

Secondly, this thesis provides added value for those with interest in the indicator questions and access to the dataset. Chapter 3 in this document contains a more or less complete analysis of how these questions relate to each other and in what manner they should be combined to form latent variables. It also describes an approach applicable for anyone with a similar dataset.

Thirdly, this thesis contains an application of the model framework for the case of HSR in Norway. This model is simple with respect to specification of utility functions; however, it sheds light on aspects important for the utility of HSR that are easily forgotten in conventional analyses. This includes in particular the heterogeneity in how individuals' utilities are affected by changes in comfort, and the "purchase of moral satisfaction" by traveling more environmentally friendly. I believe my thesis can provide a motivation for considering such latent variables in further studies in this field. I have also outlined suggestions to indicate how more sophisticated analyses should be done.

[2]I believe for instance that level of education would have increased the explanatory power for the variable "global environmental consciousness" significantly, but level of education of the respondents was not available.

References

Abbe, E., M. Bierlaire, and T. Toledo (2007). Normalization and correlation of cross-nested logit models. *Transportation Research Part B: Methodological 41*(7), 795–808.

Aigner, D., C. Hsiao, A. Kapteyn, and T. Wansbeek (1984). *Latent variable models in econometrics, chapter 23 from Handbook of econometrics*, Volume 2. (Editors: Heckman, J.J. and E.E. Leamer). Amsterdam, Elsevier.

Ajzen, I. and M. Fishbein (1980). *Understanding attitudes and predicting social behaviour*. Prentice-Hall.

Amemiya, Y. and T. Anderson (1990). Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics 18*, 1453–1463.

Anderson, T. and Y. Amemiya (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *The Annals of Statistics 16*(2), 759–771.

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *The Journal of Political Economy 97*(6), 1447–1458.

Ashok, K., W. Dillon, and S. Yuan (2002). Extending discrete choice models to incorporate attitudinal and other latent variables. *Journal of Marketing Research 39*(1), 31–46.

Atasoy, B., A. Glerum, R. Hurtubia, and M. Bierlaire (2010). Demand for public transport services: Integrating qualitative and quantitative methods. In *10th Swiss Transport Research Conference, Ascona*.

Atkins (2012a). Norway high speed rail assessment study: Phase iii – market, demand and revenue analysis. *Accessible at: http://www.jernbaneverket.no/PageFiles/ 17564/Market,_Demand_and%20Revenue_AnalyAna_Final_Report_atkins_.pdf*.

Atkins (2012b). Norway high speed rail assessment study: Phase iii – model development report. *Accessible at: http://www.jernbaneverket.no/PageFiles/17406/Economic%20and %20Financial%20Analysis%20-Final%20Report%20Atkins.pdf*.

Ben-Akiva, M., D. McFadden, T. Gärling, D. Gopinath, J. Walker, D. Bolduc, A. Börsch-Supan, P. Delquié, O. Larichev, T. Morikawa, et al. (1999). Extended framework for modeling choice behavior. *Marketing Letters 10*(3), 187–203.

Ben-Akiva, M., J. Walker, A. Bernardino, D. Gopinath, T. Morikawa, and A. Polydoropoulou (2002). Integration of choice and latent variable models. *In perpetual motion: Travel behaviour research opportunities and application challenges, (Editor: Mahmassan H.S.), Elsevier, Amsterdam*, 431–470.

Bierlaire, M. (2003). Biogeme: A free package for the estimation of discrete choice models. *Proceedings of the 3rd Swiss Transportation Research Conference*.

Bierlaire, M. (2006). A theoretical analysis of the cross-nested logit model. *Annals of operations research 144*(1), 287–300.

Bierlaire, M. and M. Fetiarison (2009). Estimation of discrete choice models: extending biogeme. In *Proceedings of the 9th Swiss Transport Research Conference. Ascona, Switzerland*.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 16*(3), 297–334.

Denstadli, J. and A. Gjerdåker (2011). Transportmiddelbruk og konkurranseflater i tre hovedkorridorer. *TØI report 1147/2011. Institute of Transport Economics, Oslo*.

Flügel, S. (2011, August). The effects of attitudinal variables on revealed current mode choice and stated future mode choice including high-speed rail options. TØI Working paper 50223. First version with comments from Patricia Mokhtarian.

Flügel, S. and A. Halse (2012). Non-linear utility specifications in mode choice models for high-speed rail. *Kuhmo Nectar Conference, Berlin*.

Flügel, S., A. Halse, J. Ortùzar, and L. Rizzi (2012). Forecasting ridership for a new mode using binary stated choice data - methodological challenges in studying the demand for high-speed rail in Norway. *1st European Symposium on Quantitative Methods in Transportation Systems, Lausanne*.

Goldberger, A. (1972). Structural equation methods in the social sciences. *Econometrica 40*(6), 979–1001.

Halse, A. (2012, July). Demand for high speed rail in norway: Survey design and data collection. TØI Working paper 50067.

Jernbaneverket (2012). Høyhastighetsutredningen 2010-2012, konklusjoner og oppsummering av arbeidet i fase 3, del 1. *Accessible at: http://www.jernbaneverket.no/PageFiles/17299/Rapport_Del_1.pdf*.

Johansson, M., T. Heldt, and P. Johansson (2006). The effects of attitudes and personality traits on mode choice. *Transportation Research Part A: Policy and Practice 40*(6), 507–525.

Johnson, R. and D. Wichern (1988). *Multivariate statistics, a practical approach*. Chapman & Hall London.

Joliffe, I. and B. Morgan (1992). Principal component analysis and exploratory factor analysis. *Statistical methods in medical research 1*(1), 69–95.

Jolliffe, I. (2005). *Principal component analysis*. John Wiley and Sons, Inc., Hoboken, New Jersey.

Jong, J. and S. Kotz (1999). On a relation between principal components and regression analysis. *The American Statistician 53*(4), 349–351.

Joreskog, K. and D. Sörbom (1977). Statistical models and methods for analysis of longitudinal data. *Latent variables in socio-economic models 103*, 285.

Kahneman, D. and J. Knetsch (1992). Valuing public goods: the purchase of moral satisfaction. *Journal of environmental economics and management 22*(1), 57–70.

Krantz Lindgren, P. (2001). *Att färdas som man lär? Om miljömedvetenhet och bilåkande*. Gidlunds förlag, Hedemora (in Swedish).

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics (Editor: Zarembka, P.)*, pp. 105–142. Academic press, New York.

McFadden, D. (1978). *Modelling the choice of residential location*. Institute of Transportation Studies, University of California.

McFadden, D. (1986). The choice theory approach to market research. *Marketing science 5*(4), 275–297.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica: Journal of the Econometric Society 57*(5), 995–1026.

McFadden, D. (1999). Rationality for economists. *Journal of risk and uncertainty 19*(1), 73–105.

McFadden, D. (2000). Disaggregate behavioral travel demand's rum side a 30-year retrospective.-presentation at the international association of travel behavior analysts. *Brisbane, Australia, July 2*, 2000.

Morikawa, T. (1989). *Incorporating stated preference data in travel demand analysis*. Ph. D. thesis, Massachusetts Institute of Technology.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika 49*(1), 115–132.

Papola, A. (2004). Some developments on the cross-nested logit model. *Transportation Research Part B: Methodological 38*(9), 833–851.

Rencher, A. and W. Christensen (2012). *Methods of multivariate analysis, Third Edition*. John Wiley and Sons, Inc., Hoboken, New Jersey.

Robinson, P. (1974). Identification, estimation and large-sample theory for regressions containing unobservable variables. *International Economic Review 15*(3), 680–692.

Rose, J., M. Bliemer, D. Hensher, and A. Collins (2008). Designing efficient stated choice experiments in the presence of reference alternatives. *Transportation Research Part B: Methodological 42*(4), 395–406.

Smith, L. (2002). A tutorial on principal components analysis. *Cornell University, USA. Discussion paper 51*, 1–52.

Sörbom, D. and K. Jöreskog (1981). The use of LISREL in sociological model building. *In Factor Analysis and Measurement in Sociological Research (Editors: D.J. Jackson and E.F. Borgatta). Beverly Hills, California: Sage Publications*, 179–199.

Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology 15*(2), 201–292.

StataCorp (2011). Stata statistical software: Release 12. *College Station, TX: StataCorp LP*.

Suhr, D. (2005). Principal component analysis vs. exploratory factor analysis. *SUGI 30 Proceedings*, 203–30.

Sydsæter, K. and B. Øksendal (1996). *Lineær algebra*. Gyldendal Norsk Forlag AS.

Train, K. (2003). *Discrete choice methods with simulation*. Cambridge University Press.

Walker, J. (2001). *Extended discrete choice models: integrated framework, flexible error structures, and latent variables*. Ph. D. thesis, Massachusetts Institute of Technology.

Appendices

A Additional descriptive analysis

A.1 Summary statistics

Table A.1: Summary statistics of indicator variables.

Indicator	Mean	S.D.	N
1	3.500	0.957	827
2	3.935	1.008	827
3	3.170	1.255	827
4	4.011	0.951	827
5	4.182	0.878	827
6	3.384	0.997	827
7	3.907	0.912	827
8	2.347	1.113	827
9	3.589	0.982	827
10	2.812	1.315	827
11	3.040	1.086	827
12	3.967	1.225	816
13	4.713	0.661	825
14	3.369	0.815	618
15	2.340	1.048	613
16	3.615	1.512	742
17	3.918	0.727	778
18	3.725	1.095	618
19	4.167	0.873	613
20	2.797	1.163	816
21	4.001	0.820	823
22	2.723	1.289	826
23	3.240	0.800	616

Table A.2: Correlation matrix of behavioral and attitudinal indicators.

$N = 502$	Indicators:																		
Indicators:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1.000																		
2	0.198	1.000																	
3	0.123	0.490	1.000																
4	0.264	0.196	0.184	1.000															
5	0.204	0.493	0.379	0.322	1.000														
6	0.193	0.297	0.307	0.263	0.471	1.000													
7	0.217	0.282	0.239	0.355	0.452	0.364	1.000												
8	0.163	-0.170	-0.071	0.062	-0.147	-0.010	0.005	1.000											
9	0.211	0.017	0.125	0.273	0.140	0.234	0.181	0.218	1.000										
10	0.090	-0.256	-0.210	0.085	-0.234	-0.051	-0.008	0.270	0.252	1.000									
11	0.233	-0.128	0.040	0.159	-0.037	0.083	0.107	0.326	0.506	0.391	1.000								
12	0.113	0.050	0.097	0.043	-0.011	0.005	0.010	-0.013	0.025	0.036	0.064	1.000							
13	0.084	-0.002	-0.004	0.051	0.007	-0.031	0.066	-0.072	0.031	0.021	0.061	0.205	1.000						
14	0.072	0.036	0.087	0.080	0.095	0.063	0.107	-0.094	0.013	0.044	-0.022	0.085	0.066	1.000					
15	0.108	0.079	0.112	0.078	0.072	0.083	0.026	0.034	0.032	0.035	0.109	0.235	0.052	0.302	1.000				
16	0.043	0.018	0.070	0.067	0.021	0.007	0.031	0.010	0.037	0.049	0.100	0.210	0.086	0.069	0.075	1.000			
17	0.092	0.041	0.006	-0.003	-0.059	-0.039	-0.054	0.094	-0.059	-0.021	0.026	0.182	0.090	-0.006	0.181	0.085	1.000		
18	0.055	-0.040	-0.049	-0.033	-0.028	0.016	-0.003	0.099	0.061	0.059	0.097	0.268	0.161	0.012	0.189	0.286	0.221	1.000	
19	0.004	-0.038	-0.064	-0.046	-0.010	-0.003	-0.055	-0.033	-0.102	-0.012	-0.044	0.152	0.165	-0.066	0.042	0.064	0.303	0.256	1.000
20	-0.025	0.170	0.038	-0.046	0.050	0.008	-0.018	0.030	-0.034	-0.063	-0.022	-0.007	0.070	0.038	0.029	0.010	0.097	0.020	0.007
21	0.073	0.109	0.080	0.002	0.056	-0.021	0.006	-0.042	-0.037	-0.094	-0.029	0.194	0.084	0.073	0.089	0.055	0.121	0.163	0.039
22	0.129	0.182	0.026	-0.015	0.052	-0.073	-0.034	-0.051	-0.122	-0.130	-0.094	0.317	0.155	0.087	0.254	0.120	0.210	0.160	0.124
23	0.068	0.018	0.027	-0.039	-0.007	0.047	0.005	-0.095	0.008	-0.025	0.021	0.225	0.099	0.112	0.149	0.066	0.160	0.190	0.112

	Indicators:			
Indicators:	20	21	22	23
20	1.000			
21	0.115	1.000		
22	0.207	0.208	1.000	
23	0.056	0.071	0.193	1.000

Note: The 23 indicators of which the correlations are displayed in this table are the same as the 23 questions regarding behaviors and attitudes. It is assumed that indicators 1–4 relate to comfort, indicators 5–7 relate to reliability, indicators 8–11 relate to flexibility, indicators 12–15 relate to local environmental consciousness, indicators 16–19 relate to safety while indicators 20–23 relate to global environmental consciousness. Horizontal and vertical lines are added to group these indicators together. The questions that are formulated with a negative meaning are reversed, so that the correlations are meaningful.

A.2 Exploratory factor analysis

Table A.3: Factor loadings and uniquenesses resulting from an EFA restricted to three factors.

$N = 502$	Factor loadings			
Indicators	1	2	3	Specificity
1	0.29	0.17	0.31	0.78
2	0.66	0.09	-0.19	0.52
3	0.57	0.05	-0.02	0.67
4	0.42	-0.01	0.31	0.73
5	0.74	-0.02	-0.02	0.45
6	0.55	-0.04	0.17	0.66
7	0.54	-0.04	0.20	0.67
8	-0.15	0.00	0.44	0.78
9	0.20	-0.05	0.60	0.60
10	-0.25	-0.02	0.52	0.66
11	-0.01	0.05	0.69	0.52
12	0.04	0.53	0.06	0.71
13	0.02	0.30	0.05	0.91
14	0.15	0.17	0.02	0.95
15	0.12	0.40	0.10	0.82
16	0.03	0.30	0.11	0.90
17	-0.05	0.44	0.00	0.81
18	-0.07	0.49	0.14	0.73
19	-0.09	0.34	-0.06	0.87
20	0.07	0.16	-0.11	0.96
21	0.09	0.30	-0.09	0.89
22	0.08	0.52	-0.17	0.69
23	0.03	0.36	-0.02	0.87

Table A.4: Predicted EFA factors.

$N = 502$	Predicted factors		
Indicators	1	2	3
1	0.07	0.06	0.11
2	0.25	0.03	-0.11
3	0.17	0.01	-0.02
4	0.11	-0.01	0.12
5	0.31	-0.03	-0.03
6	0.16	-0.03	0.07
7	0.16	-0.03	0.08
8	-0.04	0.00	0.16
9	0.05	-0.04	0.25
10	-0.08	0.00	0.21
11	-0.02	0.03	0.33
12	0.00	0.22	0.02
13	0.00	0.10	0.01
14	0.04	0.06	0.01
15	0.03	0.15	0.03
16	0.00	0.10	0.03
17	-0.02	0.17	0.00
18	-0.04	0.21	0.05
19	-0.03	0.13	-0.02
20	0.01	0.05	-0.04
21	0.02	0.10	-0.03
22	0.01	0.23	-0.07
23	0.01	0.12	-0.01

B Additional information regarding the dataset

This annex will describe this dataset in detail, in particular aspects that were not mentioned in chapter 2. It is heavily based on a working paper from Institute of Transport Economics (Halse, 2012)[1], which the reader is referred to if more information is desired. The next sections will first give information regarding questions from the RP survey, and then information regarding the questions and structure of the SP survey. Regarding the SP survey, the process of recruiting respondents, the questionnaire design and the choice experiment design will be discussed, respectively.

B.1 The revealed preference survey

The most important questions from the RP survey about the reference trip gave information regarding:

- Time and place for arrival and departure;
- Travel cost;
- Travel route;
- Reason for mode choice;
- Purpose of journey (where the most important division in this context is business or leisure);
- How frequently the trip is undertaken, and choice of mode for the previous trip;
- Number of persons travelling together;
- Age, gender, occupation, income and place of residence;
- A question about whether the respondent would be willing to be contacted at a later date for another survey regarding high speed rail.

In total, 8,450 individuals responded to the RP survey.

B.2 Recruiting respondents for the SP survey

When programming the questionnaire, an efficient design was used (see section B.4), and therefore the data collection had to consist of three rounds: two smaller pilot rounds for calibration of the design, and then the main survey. A set of factors were used for deciding which respondents that could be recruited from the RP survey to the SP survey. The most important were: respondents had given valid e-mail addresses; all reported travel attributes were defined as valid, or within the boundaries of what

[1] In particular, all lists that appear in this chapter are copied from that paper.

Table B.1: Responses for the SP survey.

Survey part:	Invited respondents:	Number of responses:	Share (%):
First pilot	826	217	26.3
Second pilot	376	67	17.8
Main study	2338	605	25.9

was expected to be reasonable; travelers had reported a positive travel cost[2]; had, if traveling by train, reported valid train stations; and had traveled between the Oslo region and either the Bergen or Trondheim region, based on the definitions below:

- *Oslo region:* The whole counties of Oslo, Akershus, Østfold and Vestfold and the municipalities Lunner, Jevnaker, Gran, Østre Toten, Venstre Toten, Søndre Land and Gjøvik in Oppland county, Eidskog, Kongsvinger, Sør-Odal, Nord-Odal, Grue, Åsnes, Stange, Våler, Hamar, Løten and Elverum in Hedmark county and Hurum, Drammen, Røyken, Kongsberg, Øvre Eiker, Nedre Eiker, Lier, Modum, Hole and Ringerike in Buskerud county.
- *Bergen region:* The whole county of Hordaland.
- *Trondheim region:* The whole county of Nord-Trøndelag, and Sør-Trøndelag County except the municipalities Røros, Tydal and Holtålen.

For the modal choice bus, only observations for the corridor Oslo-Trondheim are available; observations for the corridor Oslo-Bergen were not collected, since the number of responses was expected to be too low.

The overall response rate was difficult to calculate since some e-mail addresses were corrected multiple times. However, the share of completed questionnaires is roughly 25 percent. Considering this, and the fact that only about 40 percent of the respondents from the RP study left their e-mail addresses, there is clearly some selection bias present (Flügel, 2011). Table B.1 summarizes the number of respondents for the SP survey, for the first pilot, the second pilot and the main study respectively.

B.3 Questionnaire design for the SP survey

The most important questions for estimating the demand for high speed rail are the modal choice questions. The concept was to use the characteristics of the reference trip reported in the RP survey as input to these choices. To do this some additional questions regarding the respondents' reference trip had to be asked, in order to link the trip to a possible journey by high speed rail. The questionnaire consisted of:

- An introduction presenting the purpose of the study and the reference trip which the respondents were to recall;
- Additional questions about the reference trip;
- Questions about how the respondents would have planned the trip if they were to do it by high speed rail instead;
- The choice experiments CE1 and CE2[3];
- Control questions about choice task interpretation and choice behavior;
- Questions about how often the respondents would travel along the corridor, with and without high speed rail;

[2]This condition was relaxed in some cases to a non-negative travel cost in order to increase the sample.

[3]These will be described in the next subsection.

- Questions about travel preferences and everyday behavior, used to deduce differences in attitudes towards comfort, flexibility, reliability, safety and environmental issues.

The most important questions regarding the third bullet point was at which station respondents would board and leave if they were to make the trip by high speed rail. Those traveling from (to) Oslo to (from) Trondheim could choose between boarding (leaving) the train at Oslo Central, Gardermoen and Stange station and leaving (boarding) the train at Trondheim Central or Tynset station. Those who traveled from (to) Oslo to (from) Bergen could board (leave) the train at either Oslo Central, Lysaker, Sandvika or Hønefoss station and leave (board) the train at Bergen Central or Voss station.

B.4 Choice experiment design for the SP survey

This part of the questionnaire consisted of two different choice experiments where the attribute values of HSR varied; in choice experiment 1 (CE1) there were eight tasks (choices between the reference trip and high speed rail) where the characteristics of the reference trip were not allowed to vary, while in choice experiment 2 (CE2) there were six such tasks. In this case the attribute values of the reference trip varied with certain percentage points below and above the reference value. In CE2 there was also a third alternative, *none*, in case the respondent did not want to travel with either of the alternatives. This means that each respondent makes 15 choices in total; one RP choice assumed to be between car, train, bus and plane and 14 SP choices between the reference trip (the RP choice) and high speed rail, of which eight are contained in CE1 and six in CE2.

The experimental design used in the choice experiments was based on theory of efficient design, where the idea is that attribute level combinations presented should be generated in such a way that the most precise estimates of the utility function parameters will be obtained. The first pilot survey had an orthogonal design, meaning that attribute levels were combined randomly. The second pilot survey used an efficient design based on the results from the first pilot. The main survey used a design based on the joint estimation of the datasets from the two pilot surveys. There is a lot of information available regarding the study and the efficient design used (see Rose et al. (2008) for information regarding the efficient design in general and Halse (2012); Flügel and Halse (2012); Flügel et al. (2012) for information regarding the structure and design of this dataset).

Figure B.1 gives an example of a choice experiment from CE1. Here, one can see how the questionnaire interface looks like for the respondent. The question reads *"Which of the following modes of travel would you have chosen?"*. The six different attributes in both CE1 and CE2 are (1) total cost, (2) in-vehicle time, (3) access time, (4) egress time, (5) frequency and (6) tunnel share (percentage of travel time in tunnel) for all choice of modes.

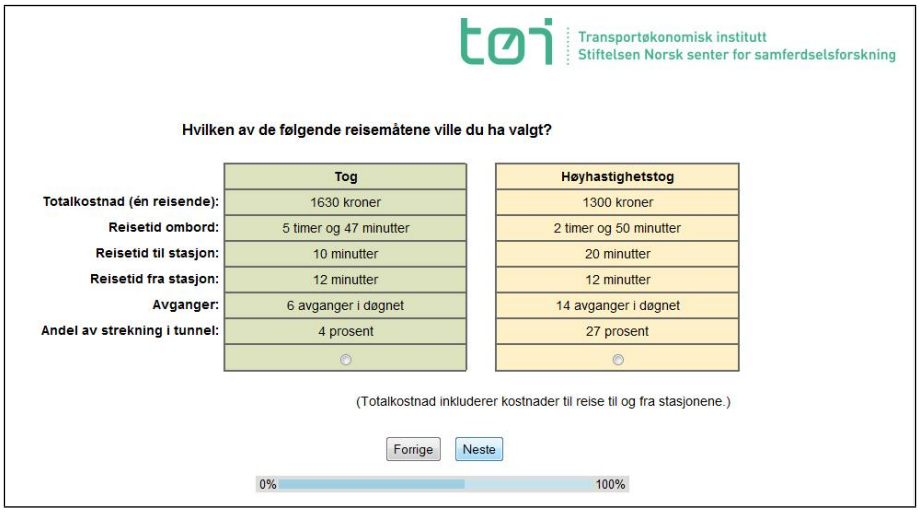

Figure B.1: Example of *choice experiment 1*, a stated choice between regular train and high speed train.

C Theoretical annex

This appendix contains sections describing theory that is relevant for the thesis but not important enough to be included directly in the document. Section C.1 describes the theory behind eigenvectors and eigenvalues. Eigenvalues are used in both principal components analysis (PCA) and the principal components approach. The principal components approach is the conventional way to estimate factor loadings in exploratory factor analysis (EFA).

Section C.2 is a short description of the theory behind PCA. I don't use PCA in this thesis at all, but the theory behind principal components analysis is crucial for understanding how factor loadings in EFA can be estimated by means of the principal component approach, which is both described and conducted in chapter 3.

Section C.3 contains a comparative discussion of PCA and EFA. My motivation for including this is that these methods seems nearly equivalent at first glance. Both methods are able to maintain most of the information from the indicator variables while reducing the dimensionality. However, I argue that PCA is inappropriate for my particular dataset, most importantly since it does not assume the existence of underlying factors or latent variables.

Section C.4 is a description of a latent variable model. More specific, it describes the MIMIC model expanded to include more than one latent variable. I also explain (shortly, in footnotes) how this model can be expanded further by including causal links between the latent variables. This is not relevant for my particular contribution; however, it is relevant for further research on the subject, described in chapter 5.

Section C.5 describes the theory behind discrete choice models in general, and binary choice models more in detail. It also contains examples of two binary choice models; the binary logit model which I use in my analysis in chapter 4 and the binary probit model which I use in appendix C.6. The binary logit model described here is combined with the latent variable model described in appendix C.4 to form the integrated choice and latent variable model that is described in chapter 4.

Finally, section C.6 contains an alternative estimation procedure than the one described in the thesis for integrated latent variable and choice models. This procedure is estimated in two steps, instead of simultaneous. It is therefore a step back in terms of efficiency. However, my motivation for including it in this appendix is that simultaneous maximum likelihood estimations have proven to take a lot of time and demand a high processor capacity when the dimensionality of the integral increases. In chapter 5 I outline how a model with all personality traits included may look like. If estimating such a model with simultaneous ML proves to be infeasible, this two-step approach may be utilized. This section is based on the framework described by Morikawa (1989). He uses a probit model for estimation, and to be able to use his formulas I chose to do so as well. However, any type of distribution for the error terms may be assumed.

C.1 Eigenvectors and eigenvalues

An eigenvector is a column vector that, if pre-multiplied with a matrix, results in the same vector multiplied with a single number, the eigenvalue. In mathematical terms;

if a number λ relates to a $(n \times n)$ matrix $\mathbf{C}$ in such a way that

$$\mathbf{Cx} = \lambda\mathbf{x} \tag{C.1}$$

where $\mathbf{x}$ is a $(n \times 1)$ vector, then $\mathbf{x}$ is an eigenvector for the matrix $\mathbf{C}$, and λ is the corresponding eigenvalue. This system of equations may be written as $(\mathbf{C} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, and for it to have a non-zero solution for $\mathbf{x}$, the coefficient matrix of $\mathbf{x}$ has to have a determinant equal to zero:

$$|\mathbf{C} - \lambda\mathbf{I}| = 0 \tag{C.2}$$

This is the characteristic equation of matrix $\mathbf{C}$, and a polynomial in λ. Solving this polynomial for λ yields the eigenvalues of $\mathbf{C}$. Having found all the eigenvalues and inserted them in equation C.1, it is now possible to solve for the corresponding eigenvectors.

If $\mathbf{C}$ is a symmetric matrix, we know that it has n eigenvectors with n corresponding eigenvalues (Sydsæter and Øksendal, 1996, p. 194). Furthermore, we also know that there exists a matrix $\mathbf{U}$ such that

$$\begin{aligned} \mathbf{U}'\mathbf{CU} &= \mathbf{D} \\ &= \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \end{aligned} \tag{C.3}$$

where $\mathbf{D}$ is a diagonal matrix with eigenvalues of $\mathbf{C}$ and the ith column in $\mathbf{U}$ is the eigenvector corresponding to the ith eigenvalue λ_i (Sydsæter and Øksendal, 1996, p. 208). The notation $\mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ or $\mathbf{D} = \text{diag}(\boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ is a vector containing all the λ_is can also be used for diagonal matrices, which is also done in this thesis.

C.2 Principal components analysis

The mathematical framework used in this section is a summary of the exposition found in Joliffe and Morgan (1992). PCA aims to replace the set of original variables $x_1, x_2, \dots x_m$ by a smaller set of variables, *principal components*, $z_1, z_2, \dots z_p$ (where p is smaller than m) that are linear combinations of the original variables

$$z_i = \alpha_{1i}x_1 + \alpha_{2i}x_2 + \dots + \alpha_{mi}x_m, \quad \forall i \tag{C.4}$$

in such a way that z_1 explains the maximum amount of variance and z_i explains the maximum amount of variance under the condition that z_i is orthogonal to $z_1, \dots z_{i-1}$. The principal components z are similar to the factors in EFA, but they do not include an error term. Equation C.4 can be written in vector notation as $z_i = \boldsymbol{\alpha}_i'\mathbf{x}$ where $\mathbf{x}$ is a column vector containing $x_1, x_2, \dots, x_m$ and $\boldsymbol{\alpha}_i'$ is the transpose of a similar column vector containing $\alpha_{1i}, \alpha_{2i}, \dots \alpha_{mi}$. The subscript i denotes which of the p principal components that is represented. Two restrictions are imposed to identify a solution:

1. To bound the variance of the principal components, the normalization constraints $\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_i = 1, \forall i$ are imposed.
2. The original variables are standardized to have unit variance. It is also conventional to give the variables zero mean, however not strictly necessary. By giving all the variables zero mean, the principal components all start at the multidimensional mean.

Since $\text{var}(z_1) = \boldsymbol{\alpha}_1'\mathbf{C}\boldsymbol{\alpha}_1$ where $\mathbf{C}$ is the correlation matrix for the standardized $x$es and $\boldsymbol{\alpha}_1$ is the (not yet identified) vector of coefficients, maximization of the Lagrangian

$$\mathcal{L} = \boldsymbol{\alpha}_1'\mathbf{C}\boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 - 1) \tag{C.5}$$

will fulfill the above requirements for z_1. The first order condition gives the equation $\mathbf{C}\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1$ which is an eigen equation (see equation C.1). Hence, λ is an eigenvalue. Pre-multiplication by $\boldsymbol{\alpha}_1'$ also shows that $\text{var}(z_1) = \lambda$, and hence, to maximize the variance λ must be as large as possible. Therefore $\text{var}(z_1) = \lambda_1$, where λ_1 denotes the largest eigenvalue of $\mathbf{C}$, and $\boldsymbol{\alpha}_1$ must be the corresponding eigenvector. This is done similarly for the rest of the principal components z_i, but under the $i-1$ orthogonality constraints $\boldsymbol{\alpha}_i'\mathbf{C}\boldsymbol{\alpha}_j = 0$, $j = 1, 2, \ldots, i-1$.

In this way, all principal components start at the multidimensional mean and are organized in such a way that the first principal component is the straight line through the mean that minimizes the square distances to all the observations, the next line is defined similarly but orthogonal to the first line, and so on. One can therefore think of PCA as an orthogonal linear transformation of the original data to a new coordinate system (a rotation around the multidimensional mean) such that the largest possible amount of variance lies on the axes (Jolliffe, 2005).

After the principal components transformation the dataset is more or less the same, only looked at from a different angle; the principal components contain the same amount of information as the original variables. The reduction in dimensionality can be achieved when one realizes that the i first principal components constitute the maximal amount of variance (i.e. the sum of the i first eigenvalues) possible to obtain from i variables. Hopefully, the last principal components will therefore have eigenvalues negligibly small, so that they can be removed from the dataset without significant loss of explanatory power. When one chooses the appropriate value of p, the number of principal components (by removing the last $m - p$ principal components from the dataset), the ratio of the original variation that is explained by the new p-dimensional dataset will be

$$\frac{\sum_{i=1}^{p} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{C.6}$$

Looking at equation C.3, we see that the method (1) eigen decompose the correlation matrix $\mathbf{C}$ to $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}'$ where $\mathbf{U}$ is a matrix where the ith column is the ith eigenvector $\boldsymbol{\alpha}_i$, and (2) erase the $(m-p)$ rows and columns with the lowest eigenvalues from $\mathbf{D}$ and the corresponding eigenvectors from $\mathbf{U}$ and $\mathbf{U}'$ will give the exact same result as the Lagrange method.

It is easy to see the relation between ordinary least squares (OLS) regression and PCA in two dimensions. Consider a dataset with two variables, y and x. If the assumed regression model is a standard linear one so that $y_i = \alpha + \beta x_i + \epsilon_i$ where the subscript denotes the ith observation pair of (x, y), we call $\hat{y_i} = \hat{\alpha} + \hat{\beta} x_i$ the variation in y_i explained by x_i in the estimated model. In the latter expression, $\hat{\alpha}$ and $\hat{\beta}$ are OLS estimates of α and β. The regression line $\hat{y}$ will lie in the x, y coordinate system along the axis defined by the first principal component. In the two dimensional case, one principal component will therefore explain exactly the same amount of variance as the $\hat{y}$ predicted from the linear regression model. If one chooses to include the second principal component as well, it will be a line starting at the base of the first principal component and expanding at a 90 degree angle. Since the number of principal components now is equal to the number of variables, all variation in the dataset is contained in the principal components. See Smith (2002) and Rencher and Christensen (2012, chapter 12) for examples of graphical plots of principal components in two dimensions and Jong and Kotz (1999) for a discussion on the relation between regression analysis and PCA.

C.3 EFA or PCA?

Based on the sections describing PCA and EFA and the chapter describing the data, it is obvious that EFA should be preferred to PCA for this case; it explicitly takes into account that the variables for attitudes and behaviors are hypothesized to be latent and influence the indicator variables. EFA and PCA are however often mistaken to be two ways of doing the same analysis, which is not always the case. It is therefore

proper with a comparative discussion which will be based on Suhr (2005). All the lists in this section are taken from her paper. One of her conclusions is:

> "Determine the appropriate statistical analysis to answer research questions a priori. It is inappropriate to run PCA and EFA with your data."

Both EFA and PCA are techniques for reducing the dimensionality of the data, and in this manner they are similar in a number of ways. Both EFA and PCA:

- reduce the number of variables;
- are only used with benefit when variables are highly correlated;
- assume a linear relationship;
- are large sample techniques where over-sampling can compensate for missing values.

There are, however, significant differences as well; while PCA is just an eigen decomposition of the correlation structure designed to maximize the variation explained given the number of principal components to retain, EFA assumes the existence of underlying factors that influence the observed variables. Based on this, the two most notable differences can be summarized as:

- In PCA, the principal components are calculated as linear combinations of the observed variables. In EFA on the other hand, the explained parts of the observed variables are estimated as linear combinations of the underlying factors;
- The relations for each observed variable in EFA include an error term, unlike the relations for each principal component in PCA where all variation is contained in the linear relationship of coefficients and observed variables.

The last bullet point has two important implications. Firstly, if the communalities in EFA are close to one, it means that the specific variances are close to zero. In this case, PCA and EFA would produce similar results. Secondly, while PCA is designed to account for the maximum amount of variation in the data, EFA only takes into account the variation that the variables have in common. In other words, when estimating factor loadings it is only the covariances or correlations that are taken into account, not the specific variances of each variable. These specific variances, ψ_i, can both contain variation not explained by the factors and measurement errors for the observed variable y_i. Since principal components are designed to contain all the variation of the dataset and not only the common variation, they are not interpretable in the same manner as factors.

C.4 Latent variable models

A CFA model is the least complicated form for a structural equation model (SEM), where latent variables μ_i are predicted by the use of indicators and theory bounded parameter constraints. The latent variable model described here is a sophistication of the CFA model, where in addition to predicting indicators the latent variables η_i[1] are assumed to be caused by exogenous covariates. This section will contain a brief description of these latent variable models, and is based mainly on Aigner et al. (1984); Robinson (1974).

The first latent model was introduced by Goldberger (1972) and contained only one latent variable. This was called a Multiple Indicators, Multiple Choices (MIMIC) model. Robinson (1974) expanded this to the more general form with m latent variables[2]. Consider a model with n individuals, p indicators and three vectors with k_0,

[1] The latent variables are denoted by η_i to separate them from the CFA factors μ_i.

[2] Note that an even more general form of equation C.8 in which factors are allowed to affect each other causally is $\boldsymbol{\eta} = \boldsymbol{\eta}\mathbf{B} + \mathbf{x}_0\boldsymbol{\Gamma}_0 + \mathbf{x}_1\boldsymbol{\Gamma}_1 + \boldsymbol{\zeta}$, see for instance Muthén (1984, p. 116). The covariance of $\mathbf{y}$ would then be somewhat different, and this is described in footnote 4. However, this form surpasses the requirements of this thesis.

k_1 and k_2 observable variables. Then the model system can be written as:

$$\mathbf{y} = \boldsymbol{\eta}\boldsymbol{\Lambda}' + \mathbf{x}_1\boldsymbol{\Upsilon}_1 + \mathbf{x}_2\boldsymbol{\Upsilon}_2 + \boldsymbol{\epsilon} \tag{C.7}$$

$$\boldsymbol{\eta} = \mathbf{x}_0\boldsymbol{\Gamma}_0 + \mathbf{x}_1\boldsymbol{\Gamma}_1 + \boldsymbol{\zeta} \tag{C.8}$$

Following the setup from Aigner et al. (1984, p. 1359) (with slightly different notation) this is the combined model for n individuals, so that $\mathbf{y}$ is the $n \times p$ matrix of indicators, $\boldsymbol{\eta}$ is the $n \times m$ matrix of latent variables and $\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_2$ are $n \times k_i$, $i = 0, 1, 2$ matrices of observed, exogenous variables. $\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$ are $n \times p$ and $n \times m$ matrices of disturbances, respectively, and $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}_0$, $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Upsilon}_1$ and $\boldsymbol{\Upsilon}_2$ are coefficient matrices.

A model with latent variables is composed by two types of linear equations; *structural equations* and *measurement equations*. The structural equations represent cause and effect relationships from exogenous variables $\mathbf{x}_0$ and $\mathbf{x}_1$ on the latent variables $\boldsymbol{\eta}$ (in this case equations C.8). The measurement equations (in this case equations C.7) model the effect of the latent variables $\boldsymbol{\eta}$ (and potential exogenous, observable variables $\mathbf{x}_1$ and $\mathbf{x}_2$) on the indicator variables $\mathbf{y}$.

The same identification problem as in CFA arises, and whether the model is identified or not can be determined in the same manner[3]. Identification is however easier to obtain in this latent variable framework since more observed variables are introduced, $\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_2$.

$\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Theta}$ is restricted to be diagonal so the same assumption as for FA has to be satisfied; the indicators cannot be correlated other than through the latent variables. The covariance of $\boldsymbol{\zeta}$ is denoted as $\mathrm{Cov}(\boldsymbol{\zeta}) = \boldsymbol{\Psi}$. There is no simultaneity in the model; the $\mathbf{x}$es determine $\mathbf{y}$ directly and via $\boldsymbol{\eta}$, so the causality is unidirectional. Therefore it suffices to consider the reduced form, which is

$$\mathbf{y} = \mathbf{x}_0\boldsymbol{\Gamma}_0\boldsymbol{\Lambda}' + \mathbf{x}_1(\boldsymbol{\Gamma}_1\boldsymbol{\Lambda}' + \boldsymbol{\Upsilon}_1) + \mathbf{x}_2\boldsymbol{\Upsilon}_2 + \boldsymbol{\epsilon} + \boldsymbol{\zeta}\boldsymbol{\Lambda} \tag{C.9}$$

Inserting for $\mathbf{y}$ from the expression above in $\mathrm{Cov}(\mathbf{y}) = \mathrm{E}(\mathbf{y}\mathbf{y}')$ conditional on $\mathbf{x}_i$, $i = 0, 1, 2$, it can be observed that each row has covariance matrix[4] $\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$. This model may be estimated provided that it is identifiable — Robinson (1974) presents a limited information estimation method. However, identified models may also be estimated with full information maximum likelihood (FIML) using the appropriate software[5].

C.5 Discrete choice models

A discrete choice refers to a situation where individuals are able to choose between a finite number of different alternatives. In the following sections, these kind of models will be referred to as choice models. For forecasting purposes the researcher is interested in the probabilities for choosing the different alternatives as functions of the exogenous variables. These probabilities are derived from a latent utility model, where individuals' perceived utilities for the different choices are unobserved, but the choices

[3] This is described in section 3.2.2.

[4] The reduced form of the model described in footnote 2 would then be $\mathbf{y} = (\mathbf{x}_0\boldsymbol{\Gamma}_0 + \mathbf{x}_1\boldsymbol{\Gamma}_1 + \boldsymbol{\zeta})(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}' + \mathbf{x}_1\boldsymbol{\Upsilon}_1 + \mathbf{x}_2\boldsymbol{\Upsilon}_2 + \boldsymbol{\epsilon}$, and the covariance of $\mathbf{y}$ given $\mathbf{x}$ would be $E(\mathbf{y}\mathbf{y}'|\mathbf{x}) = E[((\mathbf{x}_0\boldsymbol{\Gamma}_0 + \mathbf{x}_1\boldsymbol{\Gamma}_1 + \boldsymbol{\zeta})(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}' + \mathbf{x}_1\boldsymbol{\Upsilon}_1 + \mathbf{x}_2\boldsymbol{\Upsilon}_2 + \boldsymbol{\epsilon})((\mathbf{x}_0\boldsymbol{\Gamma}_0 + \mathbf{x}_1\boldsymbol{\Gamma}_1 + \boldsymbol{\zeta})(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Lambda}' + \mathbf{x}_1\boldsymbol{\Upsilon}_1 + \mathbf{x}_2\boldsymbol{\Upsilon}_2 + \boldsymbol{\epsilon})'|\mathbf{x}] = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B})'^{-1}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$. As stated earlier, this surpasses the requirements for this thesis and will not be further elaborated.

[5] The first available and most famous software program for these kind of estimations is LISREL, developed by Jöreskog and Sörbom (see for instance Joreskog and Sörbom (1977); Sörbom and Jöreskog (1981)). As described by Aigner et al. (1984, p. 1369–1371) LISREL compares the sample covariance matrix with the covariance structure obtained by the latent variable model framework's imposed restrictions, and then computes FIML estimates based on sample moments of second order and a normality assumption. At the current date latent variable and MIMIC models are estimable in a wide range of other programs as well.

are observed. The utilities for individual i (where the individual specific subscript is suppressed for notational simplicity) are written as

$$u_j = v(\mathbf{x}_j;\, \boldsymbol{\beta}_j) + \varepsilon_j \tag{C.10}$$

where u_j denotes the perceived utility for individual i for choice j, $j \in [1,\, J]$ and J denotes individual i's choice set. $\mathbf{x}_j$ denotes exogenous observable variables for alternative j; these can be either individual specific or alternative specific. $\boldsymbol{\beta}_j$ are the choice parameters that need to be estimated. The function v represents the deterministic part of the utility and ε_j is the random part of the utility, independent of $\mathbf{x}_j$. The utilities u_j are not observed, but the choice, denoted $\mathbf{d}$, is; it consists of the entities

$$d_j = \begin{cases} 1 & \text{if} \quad u_j \geq u_s; j \neq s, \forall j, s \in J \\ 0 & \text{otherwise} \end{cases} \tag{C.11}$$

where j denotes the different alternatives and the number 1 indicates which alternative that is chosen[6]. As previously mentioned, the entities of interest are the J conditional choice probabilities, denoted

$$P(d_j = 1|\mathbf{x}), \quad j \in [1,\, J] \tag{C.12}$$

For the rest of this section, the function v is assumed to be linear in $\boldsymbol{\beta}$ for simplicity.

C.5.1 Binary choice models

In the case of binary choice models the choice set consists of only two alternatives for all individuals. Therefore only one choice probability has to be estimated, $P(\mathbf{x};\boldsymbol{\beta}) = P(d = 1|\mathbf{x})$, where d is a choice indicator taking either the value 1 or the value 0 depending on the outcome of the choice. Assuming that $v(\mathbf{x};\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{x}$, the underlying latent variable model for individual i is written as

$$u = \boldsymbol{\beta}'\mathbf{x} + \varepsilon \qquad (u \text{ unobserved}, \mathbf{x} \text{ observed}) \tag{C.13}$$

$$d = \begin{cases} 1 & \text{if} \quad u > 0 \\ 0 & \text{if} \quad u \leq 0 \end{cases} \qquad (d \text{ observed}) \tag{C.14}$$

where $\mathbf{x}$ and $\boldsymbol{\beta}$ are $(k \times 1)$ vectors. ε is a continuously distributed error term independent of $\mathbf{x}$. The first equation is called a structural equation, while the second is called a measurement equation. Assuming a distribution for ε and denoting the cumulative distribution function (CDF) of that distribution as $G(\cdot)$, the choice probability can be written as

$$P(d = 1|\mathbf{x}) = P(u > 0|\mathbf{x}) = P(\varepsilon > -\boldsymbol{\beta}'\mathbf{x}|\mathbf{x}) = 1 - G(-\boldsymbol{\beta}'\mathbf{x}) = G(\boldsymbol{\beta}'\mathbf{x}) \tag{C.15}$$

where the last equality sign holds if one is assuming that the probability density function (PDF) of ε is symmetric around zero.

There are two main types og binary choice models which both were considered for the estimation in section 4.2; probit models and logit models. If $\varepsilon \sim \mathcal{N}(0, 1)$ the CDF of ε is

$$G(x) = \Phi(x) = \int_{-\infty}^{x} \phi(z)\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}z^2}\, dz \tag{C.16}$$

[6]This implies that the model is assuming that the alternative that gives the highest utility is chosen.

where Φ denotes the CDF and ϕ denotes the PDF of the normal distribution, then the choice model is called a probit model. If ε follows a Gumbel distribution, then the CDF is

$$G(x) = \Lambda(x) = \int_{-\infty}^{x} \lambda(z)\, dz = \frac{e^x}{1+e^x} \tag{C.17}$$

which is called a logit model. $\Lambda(\cdot)$ denotes the CDF of a logit distribution and $\lambda(\cdot)$ denotes the PDF. This is the conventional notation; however, to avoid confusion with the factor loadings matrices for which I use the notation $\mathbf{\Lambda}$, this is the only time Λ will be used for logit models. x in the two previous equations denotes the argument, which in our case is $v(\mathbf{x};\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{x}$.

While the variance in the probit model is 1, the variance for the logistic distribution used in a logit model is $\pi^2/3$. This leads to different scales on the coefficients, so that the logit and the probit estimates are not directly comparable; the models should be scaled so that the variances are equal first.

Looking at equation C.15, we see that the probability $P(d = 1|\mathbf{x})$ can be found directly from equation C.16 or C.17, depending on whether we assume that ε follows a normal or a logistic distribution and if we are able to estimate the βs consistently. If we also can assume that observations are independent and identically drawn from that distribution, this is easily done by means of MLE. Observing that $P(d = 0|\mathbf{x}) = 1 - P(d = 1|\mathbf{x})$, we can for individual i write

$$f(d|\mathbf{x};\boldsymbol{\beta}) = [G(\boldsymbol{\beta}'\mathbf{x})]^d [1 - G(\boldsymbol{\beta}'\mathbf{x})]^{1-d} \tag{C.18}$$

so that $f(\cdot) = G(\cdot)$ when $d = 1$ and $f(\cdot) = 1 - G(\cdot)$ when $d = 0$. By multiplying these functions for all individuals in the sample (provided that they are independently drawn from the population), the likelihood for observing the actual distribution of choices (given $\mathbf{x}$ and parameter values for $\boldsymbol{\beta}$) that is observed is calculated, $\prod_{i=1}^{N} f(d_i|\mathbf{x}_i;\boldsymbol{\beta})$. Taking the log of this eases calculations and does not matter for the result because $\max_x f(x) \Leftrightarrow \max_x \ln f(x)$. We then obtain the log likelihood function for N individuals

$$\sum_{i=1}^{N} \ell(\boldsymbol{\beta};\mathbf{x}_i) = \sum_{i=1}^{N} d_i \ln(G(\boldsymbol{\beta}'\mathbf{x}_i)) + (1 - d_i)\ln(1 - G(\boldsymbol{\beta}'\mathbf{x}_i)) \tag{C.19}$$

where $\ell(\cdot)$ denotes the log likelihood for individual i. The second order condition holds for both the logit and the probit model, and the first order condition gives the estimator of $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}_{ML_{probit}}$ or $\hat{\boldsymbol{\beta}}_{ML_{logit}}$ depending on whether $G(\cdot)$ is the normal or logistic CDF, respectively. Changes in probabilities by marginal changes in (continuous) xs can be found by

$$\frac{\delta P(d = 1|\mathbf{x})}{\delta x_h} = \frac{\delta G(\boldsymbol{\beta}'\mathbf{x})}{\delta x_h} = G'(\boldsymbol{\beta}'\mathbf{x})\beta_h \tag{C.20}$$

If ε is normally distributed, this marginal effect becomes $\phi(\boldsymbol{\beta}'\mathbf{x})\beta_h$, and if ε is logistically distributed it can be written as $G(\boldsymbol{\beta}'\mathbf{x})(1 - G(\boldsymbol{\beta}'\mathbf{x}))\beta_h = P(1 - P)\beta_h$[7].

[7] In section 4.2 I ultimately chose to estimate a logit model, and this is the reason. Acquire the relevant values for the arguments of $G'(\cdot)$ is difficult when one of the arguments is a latent variable. However, since marginal effects for logit models can be calculated by means of probabilities, which can be calculated based on the number of observed outcomes only, it is more feasible than the probit model. The main reason for why the probit model is still included in this section is that it is used in appendix C.6.

C.6 A two-step estimation procedure

This section describes an alternative estimation procedure for the integrated model. If it turns out that simultaneous ML is infeasible on a model where latent variables are included (because the dimensionality of the integral is too high relative to the available processor capacity), it is possible to do the ML estimation in two steps. This is implemented by (1) first estimating the latent model by appropriate software and (2) then include the estimated variables in the choice model in a consistent way. This is a step back compared to the estimation procedure described in the previous chapter because the estimation procedure is not fully efficient. Therefore it should only be done if other options are infeasible.

The method will be illustrated by means of two choice model examples; section C.6.1 contains an elegant model formulation developed and described by Morikawa (1989); McFadden (2000), while section C.6.2 contains an extension of the aforementioned method to a multinomial case, described in Johansson et al. (2006).

C.6.1 The case of a binary probit model

This section contains (1) a description of a binary probit model[8] comparable to the model described by equations 4.1–4.4 where the functions $\mathbf{h}$, $\mathbf{v}$ and $\mathbf{g}$ are linear, and (2) a consistent two-step estimation procedure for the model. This particular model was first developed by McFadden and Morikawa, used in Morikawa (1989) and summarized in McFadden (2000). The model framework described below is, with some minor deviations, a summary of Morikawa (1989), although the notation mainly is based on Johansson et al. (2006) for consistency reasons.

The model equations, corresponding to 4.1–4.4, are assumed to have the form (for individual i, suppressing individual specific subscripts)

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\mathbf{x}_1 + \boldsymbol{\zeta} \tag{C.21}$$

$$u = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{\eta} + \boldsymbol{\beta}_2'\mathbf{x}_2 + \varepsilon \tag{C.22}$$

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\xi} \tag{C.23}$$

$$d = \begin{cases} 1 & \text{if} \quad u > 0 \\ 0 & \text{if} \quad u \leq 0 \end{cases} \tag{C.24}$$

where we also assume $\boldsymbol{\zeta} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Psi})$, $\varepsilon \sim \mathcal{N}(0,1)$, $\boldsymbol{\xi} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Theta})$ and $E(\boldsymbol{\zeta}, \varepsilon) = E(\boldsymbol{\zeta}, \boldsymbol{\xi}) = E(\varepsilon, \boldsymbol{\xi}) = \mathbf{0}$. $\boldsymbol{\eta}$ is the vector of m latent, unobservable variables, $\mathbf{y}$ is the vector of p observable indicators, $\mathbf{x}_1$ is the vector of k_1 observable, exogenous variables influencing $\boldsymbol{\eta}$, $\mathbf{x}_2$ is the vector of k_2 observable, exogenous variables influencing u and $\boldsymbol{\zeta}$, ε and $\boldsymbol{\xi}$ are $(m \times 1)$, (1×1) and $(p \times 1)$ vectors of error terms, respectively; all are assumed normally distributed with mean zero. $\mathbf{x}_1$ and $\mathbf{x}_2$ may contain some of the same exogenous variables. $\mathcal{N}_q$ denotes a q dimensional multivariate normal distribution. β_0 is an intercept and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $(m \times 1)$ and $(k_2 \times 1)$ vectors of coefficients, respectively, and $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ are $(m \times k_1)$ and $(p \times m)$ coefficient matrices, respectively.

The joint multivariate normal distribution for $\mathbf{y}$, $\boldsymbol{\eta}$ and u conditional on $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ is then:

$$\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\eta} \\ u \end{pmatrix} = \mathcal{N}_{p+m+1}(\mathbf{M}_1, \boldsymbol{\Omega}_1) \tag{C.25}$$

[8] Using the probit model is only done for illustration purposes; changing the distribution of the error terms to a non-normal distribution is straight forward.

where

$$\mathbf{M}_1 = \begin{pmatrix} \boldsymbol{\Lambda\Gamma x}_1 \\ \boldsymbol{\Gamma x}_1 \\ \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{\Gamma x}_1 + \boldsymbol{\beta}_2'\mathbf{x}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Omega}_1 = \begin{pmatrix} \boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta} & \boldsymbol{\Lambda\Psi} & \boldsymbol{\Lambda\Psi\beta}_1 \\ \boldsymbol{\Psi\Lambda}' & \boldsymbol{\Psi} & \boldsymbol{\Psi\beta}_1 \\ \boldsymbol{\beta}_1'\boldsymbol{\Psi\Lambda}' & \boldsymbol{\beta}_1'\boldsymbol{\Psi} & 1 + \boldsymbol{\beta}_1'\boldsymbol{\Psi\beta}_1 \end{pmatrix} \tag{C.26}$$

The conditional distribution of $\boldsymbol{\eta}$ and u given $\mathbf{y}$ and $\mathbf{x}$ can then be deduced; see Johnson and Wichern (1988) for a proof[9]:

$$\begin{pmatrix} \boldsymbol{\eta} \\ u \end{pmatrix} = \mathcal{N}_{m+1}(\mathbf{M}_2, \boldsymbol{\Omega}_2) \tag{C.27}$$

where

$$\mathbf{M}_2 = \begin{pmatrix} \boldsymbol{\Gamma x}_1 + \boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}(\mathbf{y} - \boldsymbol{\Lambda\Gamma x}_1) \\ \beta_0 + \boldsymbol{\beta}_2'\mathbf{x}_2 + \boldsymbol{\beta}_1'\{\boldsymbol{\Gamma x}_1 + \boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}(\mathbf{y} - \boldsymbol{\Lambda\Gamma x}_1)\} \end{pmatrix} \tag{C.28}$$

and

$$\boldsymbol{\Omega}_2 = \begin{pmatrix} \boldsymbol{\Psi} - \boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}\boldsymbol{\Lambda\Psi} & \boldsymbol{\Psi\beta}_1 - \boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}\boldsymbol{\Lambda\Psi\beta}_1 \\ \boldsymbol{\beta}_1'\boldsymbol{\Psi} - \boldsymbol{\beta}_1'\boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}\boldsymbol{\Lambda\Psi} & 1 + \boldsymbol{\beta}_1'\boldsymbol{\Psi\beta}_1 - \boldsymbol{\beta}_1'\boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}\boldsymbol{\Lambda\Psi\beta}_1 \end{pmatrix} \tag{C.29}$$

The choice model, given $\mathbf{y}$ and $\mathbf{x}$, is therefore (see equation C.18)

$$\begin{aligned} P(d|\mathbf{y},\mathbf{x}) = {} & \Phi\left(\frac{\beta_0 + \boldsymbol{\beta}_2'\mathbf{x}_2 + \boldsymbol{\beta}_1'\{\boldsymbol{\Gamma x}_1 + \boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}(\mathbf{y} - \boldsymbol{\Lambda\Gamma x}_1)\}}{\sqrt{1 + \boldsymbol{\beta}_1'\boldsymbol{\Psi\beta}_1 - \boldsymbol{\beta}_1'\boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}\boldsymbol{\Lambda\Psi\beta}_1}}\right)^d \\ & \times \left(1 - \Phi\left(\frac{\beta_0 + \boldsymbol{\beta}_2'\mathbf{x}_2 + \boldsymbol{\beta}_1'\{\boldsymbol{\Gamma x}_1 + \boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}(\mathbf{y} - \boldsymbol{\Lambda\Gamma x}_1)\}}{\sqrt{1 + \boldsymbol{\beta}_1'\boldsymbol{\Psi\beta}_1 - \boldsymbol{\beta}_1'\boldsymbol{\Psi\Lambda}'[\boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}]^{-1}\boldsymbol{\Lambda\Psi\beta}_1}}\right)\right)^{1-d} \end{aligned} \tag{C.30}$$

where Φ denotes the CDF of the normal distribution. Two-step estimation consists of (1) using appropriate SEM software to estimate $\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\Theta}}$ from equations C.21 and C.23 and then calculate the fitted values

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Gamma}}\mathbf{x}_1 + \hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}}'[\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Theta}}]^{-1}(\mathbf{y} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Gamma}}\mathbf{x}_1) \tag{C.31}$$

$$\hat{\boldsymbol{\omega}} = \hat{\boldsymbol{\Psi}} - \hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}}'[\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Theta}}]^{-1}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Psi}} \tag{C.32}$$

where $\boldsymbol{\omega}$ denotes the covariance matrix of $\boldsymbol{\eta}$, and (2) use MLE to maximize the log likelihood function for the population of N individuals to obtain the choice parameters from equation C.30. Let $\ell(\cdot)$ denote individual i's log likelihood function. Then ML parameters are obtained by (analogous to equation C.19)

$$\begin{aligned} & \max_{\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \sum_{i=1}^{N} \ell(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2; d_i, \mathbf{x}_{2i}, \hat{\boldsymbol{\eta}}_i, \hat{\boldsymbol{\omega}}) = \\ & \max_{\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \sum_{i=1}^{N} \left[d_i \ln \Phi\left(\frac{\beta_0 + \boldsymbol{\beta}_1'\hat{\boldsymbol{\eta}}_i + \boldsymbol{\beta}_2'\mathbf{x}_{2i}}{\sqrt{1 + \boldsymbol{\beta}_1'\hat{\boldsymbol{\omega}}\boldsymbol{\beta}_1}}\right) + (1 - d_i) \ln \left(1 - \Phi\left(\frac{\beta_0 + \boldsymbol{\beta}_1'\hat{\boldsymbol{\eta}}_i + \boldsymbol{\beta}_2'\mathbf{x}_{2i}}{\sqrt{1 + \boldsymbol{\beta}_1'\hat{\boldsymbol{\omega}}\boldsymbol{\beta}_1}}\right)\right)\right] \end{aligned} \tag{C.33}$$

which gives the estimates $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. These estimates are consistent, but since the arguments $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\omega}}$ in the likelihood function are estimated a correction of the covariance matrix for the second step estimates is needed, see Morikawa (1989, p. 129) and McFadden (1989).

[9] Johnson and Wichern are referred to in Morikawa (1989, p. 128). I did not have access to Johnson and Wichern (1988) while writing this thesis, therefore I rely on the result referred to by Morikawa.

C.6.2 The case of a multinomial probit model

Extending this to the multinomial case is straightforward; see for instance Johansson et al. (2006); instead of the scalar u we would have had the J dimensional vector of utilities $\mathbf{u}$ with error terms $\boldsymbol{\varepsilon} \sim \mathcal{N}_J(\mathbf{0}, \boldsymbol{\Xi})$, where $\boldsymbol{\Xi} = \mathbf{I}$ (a J dimensional identity matrix) must be imposed for identification purposes. The joint multivariate normal distribution from equation C.25 would then have been $(p+m+J)$ dimensional and the term on the third row and third column of $\boldsymbol{\Omega}_1$ would have been $\boldsymbol{\Xi} + \boldsymbol{\beta}_1' \boldsymbol{\Psi} \boldsymbol{\beta}_1$ instead of $1 + \boldsymbol{\beta}_1' \boldsymbol{\Psi} \boldsymbol{\beta}_1$. Similarly, the joint multivariate normal distribution from equation C.27 would have been $(m + J)$ dimensional and the term on the second row and second column of $\boldsymbol{\Omega}_2$ would have been $\boldsymbol{\Xi} + \boldsymbol{\beta}_1' \boldsymbol{\Psi} \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1' \boldsymbol{\Psi} \boldsymbol{\Lambda}' [\boldsymbol{\Lambda} \boldsymbol{\Psi} \boldsymbol{\Lambda}' + \boldsymbol{\Theta}]^{-1} \boldsymbol{\Lambda} \boldsymbol{\Psi} \boldsymbol{\beta}_1$ instead of $1 + \boldsymbol{\beta}_1' \boldsymbol{\Psi} \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1' \boldsymbol{\Psi} \boldsymbol{\Lambda}' [\boldsymbol{\Lambda} \boldsymbol{\Psi} \boldsymbol{\Lambda}' + \boldsymbol{\Theta}]^{-1} \boldsymbol{\Lambda} \boldsymbol{\Psi} \boldsymbol{\beta}_1$. If d_j is defined as

$$d_j = \begin{cases} 1 & \text{if} \quad u_j \geq u_s; \forall j,\ s \in J \\ 0 & \text{otherwise} \end{cases} \tag{C.34}$$

then the log likelihood function of the whole sample of N individuals can be estimated by

$$\max_{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \sum_{i=1}^{N} \sum_{j=1}^{J} d_{ij} \ln \Phi \left(\frac{\beta_{0j} + \boldsymbol{\beta}_{1j}' \hat{\boldsymbol{\eta}}_i + \boldsymbol{\beta}_{2j}' \mathbf{x}_{2i}}{\sqrt{1 + \boldsymbol{\beta}_{1j}' \hat{\boldsymbol{\omega}} \boldsymbol{\beta}_{1j}}} \right) \tag{C.35}$$

analogous to equation C.33.